Vanilla Bayesian Optimization Performs Great in High Dimensions

Carl Hvarfner [1] **Erik O. Hellsten** [1 2] **Luigi Nardi** [1 2]

Abstract

High-dimensional problems have long been considered the Achilles' heel of Bayesian optimization. Spurred by the curse of dimensionality, a large collection of algorithms aim to make it more performant in this setting, commonly by imposing various simplifying assumptions on the objective. In this paper, we identify the degeneracies that make vanilla Bayesian optimization poorly suited to high-dimensional tasks, and further show how existing algorithms address these degeneracies through the lens of lowering the model complexity. Moreover, we propose an enhancement to the prior assumptions that are typical to vanilla Bayesian optimization, which reduces the complexity to manageable levels without imposing structural restrictions on the objective. Our modification - a simple scaling of the Gaussian process lengthscale prior with the dimensionality - reveals that standard Bayesian optimization works drastically better than previously thought in high dimensions, clearly outperforming existing state-of-the-art algorithms on multiple commonly considered real-world high-dimensional tasks.

1. Introduction

In Bayesian optimization, *complexity* and *dimensionality* are intrinsically interlinked — the higher the problem dimensionality, the harder it is to optimize. The exuberance of space, and large distance between observations, makes the size of high-variance regions along the boundary of the search space exponentially large (Malu et al., 2021; Binois & Wycoff, 2022). Moreover, the growing number of parameters of the Gaussian Process (GP) surrogate in relation to the number of observations makes accurate modeling of the problem at hand exceedingly difficult. In recent years, the effort to create methods that achieve efficient Bayesian optimization (BO) in high dimensions has been substantial, making it one of the most frequently addressed challenges in the BO research community (Kandasamy et al., 2015; Wang et al., 2016; Nayebi et al., 2019; Eriksson et al., 2019; Eriksson & Jankowiak, 2021; Papenmeier et al., 2022; 2023; Ziomek & Bou Ammar, 2023).

While approaches are plentiful and diverse, they all share a common characteristic: they employ restrictions on the objective which reduces its a-priori *assumed complexity* by contracting the search space. This in turn decreases distances between data points and prospective queries, increasing their correlation, thus making GP inference more informative. Assuming a degree of complexity which enables meaningful correlation is essential to efficiently optimize problems of *any* dimensionality. Nevertheless, the high-complexity, low-correlation issue presents itself most clearly in the high-dimensional setting.

In this paper, we hypothesize that the shortcomings of vanilla BO in high dimensions are strictly a consequence of the complexity assumptions imposed on the objective. To that end, we view existing high-dimensional BO (HDBO) approaches through the lens of model complexity, which arises from their structural assumptions. Thereafter, we modify standard BO to follow a similarly complexity reduced structure, simply by appropriately scaling the lengthscale prior of the GP kernel. Consequently, we effectively circumvent the well-established Curse of Dimensionality (CoD) without introducing any of the conventional structural restrictions on the objective that are prevalent in HDBO. We demonstrate that standard BO works drastically better than previously thought for high-dimensional tasks, outclassing existing high-dimensional BO algorithms on a wide range of real-world problems. Further, we aim to shed light on the inner workings of the BO machinery and why minimal changes in assumptions yield a dramatic increase in performance. The result is a performant *vanilla* BO algorithm for dimensionalities well into the thousands.

Formally, we make the following contributions:

1. We demonstrate the crucial difference between dimensionality and complexity in BO, highlighting the failure modes related to high *assumed complexity* and relate existing HDBO classes to a reduction in complexity.

[1]Lund University, Lund, Sweden [2]DBTune, Malmö, Sweden. Correspondence to: Carl Hvarfner <carl.hvarfner@cs.lth.se>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

2. We prove that when the model is uninformed, EI will *not* exhibit exploratory behavior along the boundary, contrasting claims of (Swersky, 2017; Oh et al., 2018).

3. We propose a plug-and-play enhancement to the vanilla BO algorithm that reduces the assumed complexity to enable high-dimensional optimization, and extensively validate it across a wide spectrum of dimensionalities. Results show that vanilla BO works significantly better for high-dimensional problems than previously imagined, substantially outperforming state-of-the-art HDBO methods on a wide range of real-world tasks.

2. Background

In this section, we review the background related to Gaussian processes and Bayesian optimization. We outline the maximal information gain (MIG) as a measure of problem complexity, and the model-level choices that impact the a-priori assumed problem complexity, to subsequently explore pitfalls of vanilla BO for high-complexity tasks in Sec 4.

2.1. Gaussian Processes

The Gaussian process (GP) has become the model class of choice in most BO applications. The GP provides a distribution over functions $\hat{f} \sim \mathcal{GP}(m(\cdot), k(\cdot,\cdot))$ fully defined by the mean function $m(\cdot)$ and the covariance function $k(\cdot,\cdot)$. Under this distribution, the value of the function $\hat{f}(\boldsymbol{x})$, at a given location $\boldsymbol{x}$, is normally distributed with a closed-form solution for the mean $\mu(\boldsymbol{x})$ and variance $\sigma^2(\boldsymbol{x})$. We model a constant mean, so that the dynamics are fully determined by the covariance function $k(\cdot,\cdot)$.

To account for differences in variable importance, each dimension is individually scaled using lengthscale hyperparameters ℓ_i. This is commonly referred to as Automatic Relevance Determination (ARD) (Williams & Rasmussen, 1995). For D-dimensional inputs $\boldsymbol{x}$ and $\boldsymbol{x}'$, the distance $r(\boldsymbol{x}, \boldsymbol{x}')$ is subsequently computed as $r^2 = \sum_{i=1}^{D}(x_i - x_i')^2/\ell_i^2$. Along with the signal variance σ_f and *noise variance* σ_ε^2, $\boldsymbol{\theta} = \{\boldsymbol{\ell}, \sigma_\varepsilon^2, \sigma_f^2\}$ comprise the set of hyperparameters that are conventionally learned, with a possible addition of a learnable constant mean c (Balandat et al., 2020; Snoek et al., 2012a). The likelihood surface $p(\boldsymbol{\theta}|\mathcal{D})$ for the GP hyperparameters is typically highly multi-modal (Rasmussen & Williams, 2006; Yao et al., 2020) and desirable hyperparameters are conventionally found by MAP estimation, where a hyperprior is set on the kernel hyperparameters $\boldsymbol{\theta}$. While often overlooked, the choice of hyperprior can greatly impact the performance of a BO algorithm in practice, particularly in non-conventional problem settings (Eriksson & Jankowiak, 2021; Baptista & Poloczek, 2018; Rothfuss et al., 2021a;b; Hvarfner et al., 2023).

2.2. Bayesian Optimization

We aim to find a maximizer $\boldsymbol{x}^* \in \arg\max_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ of the black-box function $f(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$, over the D-dimensional input space $\mathcal{X} = [0,1]^D$. We assume that f can only be observed point-wise and that the observations are perturbed by Gaussian noise, $y(\boldsymbol{x}) = f(\boldsymbol{x}) + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

The *acquisition function* uses the surrogate model to quantify the utility of a point in the search space. Acquisition functions employ a trade-off between exploration and exploitation, typically using a greedy heuristic to do so. Most common is the Expected Improvement (EI) (Jones et al., 1998; Bull, 2011) and its numerically stable, easy-to-optimize adaptation LogEI (Ament et al., 2023). Another acquisition function which uses similar heuristics is the Upper Confidence Bound (Srinivas et al., 2012; 2010).

A Working Definition of "Vanilla" BO

In addition to the two main components — the probabilistic surrogate model and the acquisition function - BO entails multiple hidden design choices, that are paramount to its efficiency. We consider the vanilla BO algorithm to standardize the output values, and to use either a Squared Exponential (Radial Basis Function, RBF) (Jones, 2001) or a $\frac{5}{2}$-Matérn (Snoek et al., 2012a; 2014) Kernel, with ARD lengthscales, an EI-family acquisition function and multi-start gradient-based acquisition function optimization.

While both MLE and MAP are commonly used for hyperparameter selection in practical Bayesian optimization, prevalent BO frameworks (Balandat et al., 2020; GPyOpt-authors, 2016; Hutter et al., 2011) employ MAP estimation, setting a prior on $p(\boldsymbol{\theta})$. While not included in our definition of the vanilla algorithm, the prior $p(\boldsymbol{\ell})$ commonly places high density on low values of $\boldsymbol{\ell}$. Furthermore, broad, uninformative priors are conventionally used on σ_ε^2 and σ_f^2. While fully Bayesian hyperparameter treatment (Osborne, 2010; Snoek et al., 2012a) may also be used, we do not consider it part of the vanilla algorithm.

2.3. The Maximal Information Gain

Our work centers around the *assumed* complexity of a problem, which conceptually could be seen as the size of the space of functions that have non-negligible probability under the GP prior. To quantify the assumed complexity, we use the *Maximal Information Gain* (MIG) (Srinivas et al., 2010) measure, which is the maximum obtainable information about the function from querying a fixed number of points. Firstly, we recall the *Information Gain* (IG) for a GP model and a set of points $\boldsymbol{X}$ is defined as

$$I(y_{\boldsymbol{X}}, f_{\boldsymbol{X}}) = \frac{1}{2}\log|\boldsymbol{I} + \sigma_\varepsilon^{-2}\mathbf{K}|, \tag{1}$$

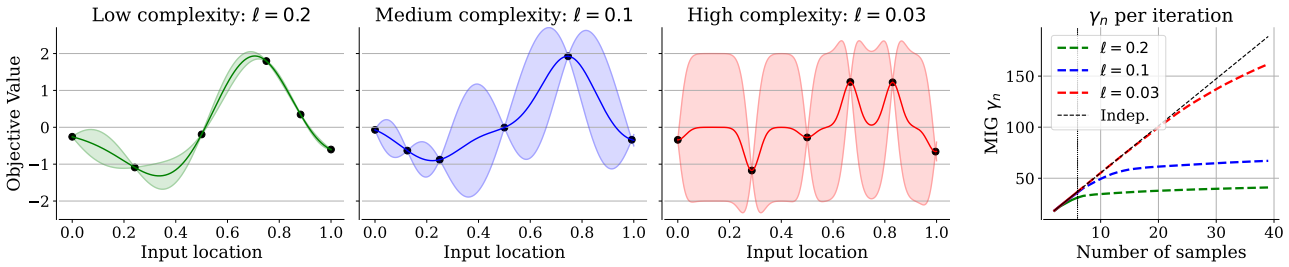


Figure 1. Three models (green, blue, red) with varying lengthscales, and thus varying complexity, attempting to model the same objective, acquiring data by greedily maximizing the IG. The MIG is shown for the three models as well as an independent kernel (dashed black), where the matrix $\mathbf{K} = \boldsymbol{I}$. The MIG for the complex model closely follows the independent kernel for 20 samples, suggesting that the complex model can acquire 20 data of approximately maximal variance. The vertical line in the MIG-plot indicates the current iteration.

where $\mathbf{K} = k(\boldsymbol{X}, \boldsymbol{X})$ is the Gram matrix for $\boldsymbol{X}$. Then, for a fixed number of points $|\boldsymbol{X}_n| = n$, the MIG is the maximizer of this measure

$$\gamma_n = \max_{\boldsymbol{X}_n \subset \mathcal{X}} I(y_{\boldsymbol{X}_n}, f_{\boldsymbol{X}_n}). \tag{2}$$

For fixed observation noise, the MIG is fully defined by the covariance matrix, which in turn depends on the choice of kernel, the problem dimensionality and the kernel hyperparameters. The MIG is maximal when the samples are independent, i.e., $\mathbf{K} \approx \boldsymbol{I}$. The MIG lacks a closed form solution, but is approximated to $(1 - 1/e)$-accuracy by sequentially querying the set of points with maximal posterior variance (Nemhauser et al., 1978; Krause et al., 2008).

Since the MIG measures the assumed complexity of f, it effectively quantifies the difficulty of optimizing a given task within a Bayesian optimization context (Srinivas et al., 2010; 2012; Berkenkamp et al., 2019), given that *the assumptions on k are accurate*. As long as the the MIG is nearly linear in the number of observations, there are regions in the search space that are almost independent of the collected data under the model. As such, there are still locations that we know nothing about, which makes optimization difficult. On the contrary, a small growth rate of the MIG suggests that the model would learn little by querying an additional point.

In Fig. 1, we provide an intuition for the MIG. We show a simpler model (left, green), as well as increasingly complex models (blue, red) for six data points. Their associated MIGs (right) are displayed for the current iteration (solid) and subsequent iterations (dashed) up until iteration 40. For the simpler model, there is little left to learn about the function, and as such, its subsequent MIG growth is small. *If* the green model is accurate, subsequent optimization is trivial due to efficient modelling that stems from large correlation in the data. On the contrary, the almost-zero correlation displayed in the red model makes its optimization vastly more difficult. This point is further emphasized by its MIG, which starts to deviate substantially from an independent kernel first after 20 data points. This suggests that *the model has capacity for 20 almost-maximal variance data acquisitions*, despite modelling only a one-dimensional objective.

3. Related Work

Multiple approaches have been proposed to tackle the limitations of BO in high dimensions. These resort to structural assumptions on the objective, which we outline by class.

Low-dimensional active subspaces
Subspace methods assume the existence of a lower-dimensional space which is representative of the function in the full-dimensional space. The active subspace can can be either axis-aligned (Nayebi et al., 2019; Eriksson & Jankowiak, 2021; Papenmeier et al., 2022; 2023) or non-axis-aligned (Garnett et al., 2014; Wang et al., 2016; Kirschner et al., 2019; Binois et al., 2020; Letham et al., 2020). Explicit variable selection approaches (Djolonga et al., 2013; Li et al., 2017; Wang et al., 2020; Song et al., 2022; Hellsten et al., 2023) employ the axis-aligned assumption to identify important variables to optimize over.

Additive kernels
AddGP methods (Duvenaud et al., 2011; Kandasamy et al., 2015; Gardner et al., 2017; Wang et al., 2017; Rolland et al., 2018; Han et al., 2021; Ziomek & Bou Ammar, 2023) decompose the objective into a sum of low-dimensional component functions, where by assumption each component is only impacted by a small subset of all variables. As such, the maximal dimensionality of each component is substantially lower than the full dimensionality of f.

Local Bayesian optimization
These approaches (Eriksson et al., 2019; Müller et al., 2021; Wan et al., 2021; Nguyen et al., 2022; Wu et al., 2023) adaptively restrict the search space to combat the CoD, limiting the optimization to a subset of the search space. By focusing on a smaller portion of the search space, the model exhibits less variation than a global model, which simplifies opti-

mization. Moreover, enforcing local optimization decreases the susceptibility of the optimizer to the model.

Non-Euclidean kernels

These methods are employed to escape the exponential growth of the typical hypercube search space in the dimensionality of the problem. Cylindrical kernels (Swersky et al., 2014; Oh et al., 2018) transform the geometry of the search space, which consequently expands the center of the search space, shrinking the boundaries.

The three pieces of related work that are most similar to ours are Elastic GPs (Rana et al., 2017), SAASBO (Eriksson & Jankowiak, 2021) and BOCK (Oh et al., 2018), which all perform optimization in the full-dimensional search space. Of these, (Rana et al., 2017) consider various lengthscales of the GP when optimizing the acquisition function, but uniquely does not impose simplifying assumptions on the model. (Eriksson & Jankowiak, 2021) and BOCK (Oh et al., 2018) employ their aforementioned assumptions to facilitate effective optimization. Contrary to these works, we facilitate optimization in the ambient space without making any of the specific structural assumptions outlined in Sec. 3.

4. Pitfalls of High-Complexity Assumptions

We now discuss the issues related to highly complex models and connect it to the high-dimensional setting. Sec. 4.1 demonstrates the intuitive relation between complexity and dimensionality. Building upon the intuitive understanding of the MIG and the related model design choices gained in Sec. 2.3, we delve into the BO-specific pathology that arises from an overly complex model in Sec. 4.2, proving that it is distinct from the well-known *boundary issue* (Swersky, 2017). Thereafter, we demonstrate how various HDBO methods circumvent the high-complexity-issue by showing the how conventional structural assumptions reduce the model complexity. In subsequent sections, we will use the terms MIG and complexity interchangeably.

4.1. Complexity and Dimensionality

Increased model complexity most often becomes a critical issue for BO algorithms in high-dimensional problems — with increasing dimensions, the maximal space between points increases. Specifically, the expected distance between randomly sampled points in a unit cube increases proportional to the square root of the dimension (Köppen, 2000). For both the RBF and the Matern-$\frac{5}{2}$ kernels, this greatly impacts the covariance, which decreases exponentially with the $\boldsymbol{\ell}$-normalized squared distance.

In Fig. 2, we display the scaling of the complexity in the number of data points, considering an RBF-kernel with fixed lengthscales. The curves represent increasing dimensionalities. As the dimensionality increases, the covariance matrix increasingly resembles that of an independent kernel (black dashed line). For $D = 18$ (purple), this manifests in a visible difference first after 3000 samples, whereas for $D = 24$ (yellow), 5000 samples is insufficient to distinguish an RBF kernel from an independent one, which implies that $k(\boldsymbol{X}_{5000}, \boldsymbol{X}_{5000}) \approx \boldsymbol{I}$. This strongly suggests that global modelling of the objective is uninformative, as the model quickly reverts back to its prior mean and variance even after collecting vast amounts of data, and meaningful inference between observed data points becomes very difficult.

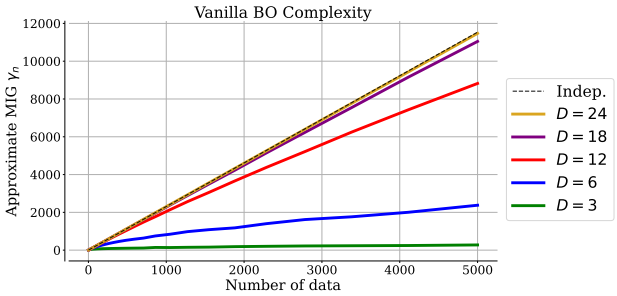
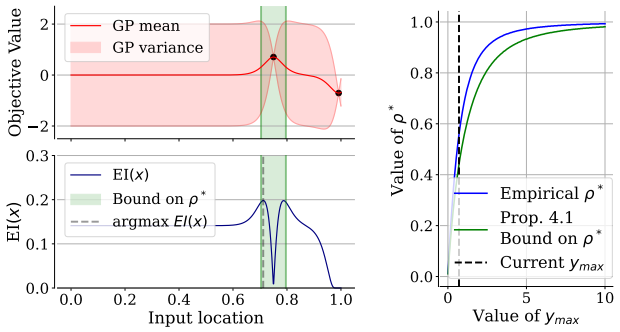


Figure 2. Complexity scaling in the number of data points for varying dimensionalities of the problem for vanilla BO with a lengthscale of $\boldsymbol{\ell} = 0.5$. For $D = 18$, the complexity visually differs from an independent kernel after approximately 3000 data points. For $D = 24$, 5000 data points are not sufficient to rid independence between observations. The MIG is approximated by sampling evenly distributed data using a SOBOL sequence.

4.2. The Boundary Issue Revisited

As covered in Sec. 2.3, high complexity implies that k produces relatively low correlation both between acquired data points and in prospective queries. Thus, the GP will only be informative close to existing observations, and will quickly revert back to the prior as we move away from this data, as demonstrated in the rightmost model in Fig. 1. Under this regime, the BO data acquisition will be highly dependent on the hyperparameters that dictate said GP prior, namely the signal variance σ_f^2 and the mean constant c. We note that, while these parameters are not always learned (De Ath et al., 2020), the choice to fix them to 1.0 and 0.0 respectively influences on the behavior of the BO algorithm as well.

Introduced by Swersky (2017), the boundary issue is the phenomenon that `EI` will, in high-dimensional settings, repeatedly query uninformed, high-variance points along the boundary of the search space to maximally explore in light of an uninformed model. We contrast this claim by the following proposition, which demonstrates that `EI` does *not* tend towards maximal variance when the model is uninformed, namely when $\mathbf{K} \approx \boldsymbol{I}$ (as in the $D = 24$ example in Fig. 2). We denote by $\boldsymbol{x}_{\text{inc}}$ the location of the incumbent, its value by $y_{\max}$, and the GP mean function by c.



a) **b)**

Figure 3. Lower bound on the optimal correlation ρ^* between the incumbent and the upcoming query. **a)** The GP for two almost-independent observations with a large exploratory region. EI prefers to query close to the incumbent, well within the bound on ρ^* from Prop. 4.1. **b)** Tightness of the bound compared a numerical solve for optimal correlation for various values of y_{max}.

Proposition 4.1 (Lower Bound on EI Correlation). *Assume that $y_{max} > c$, $\mathbf{K} = \sigma_f^2 \boldsymbol{I}$ and that the candidate query $\boldsymbol{x}_*$ correlates with at most one observation. Then, the correlation $\rho^* = \sigma_f^{-2} k(\boldsymbol{x}_*, \boldsymbol{x}_{inc})$ between the next query $\boldsymbol{x}_* = \arg\max_{\boldsymbol{x} \in \mathcal{X}} \mathtt{EI}(\boldsymbol{x})$ and $\boldsymbol{x}_{inc}$ satisfies*

$$\rho^* \sqrt{\frac{1+\rho^*}{1-\rho^*}} \geqslant \frac{y_{max} - c}{\sigma_f}. \tag{3}$$

We outline the proof in App. C, where we parameterize EI by the correlation ρ between a candidate $\boldsymbol{x}_*$ and the incumbent, and show that $\frac{\partial \mathtt{EI}}{\partial \rho}$ is positive for all values of ρ below the bound in Eq. 3. Thus, EI will prefer an observation that has substantial correlation with $\boldsymbol{x}_{inc}$ to one that does not.

Proposition 4.1 demonstrates that, when correlation in the model is low, EI does not seek out high-variance regions as described by Swersky (2017). This contrasts the common belief that HDBO intrinsically suffers from excessive exploration around the borders (Siivola et al., 2017; Oh et al., 2018; Eriksson et al., 2019; Eriksson & Jankowiak, 2021; Müller et al., 2021; Binois & Wycoff, 2022; Eduardo & Gutmann, 2022). As we will soon cover, uninformative models frequently display the opposite behavior, where repeated queries are made exceedingly close to the incumbent.

In Fig. 3a, we observe this query behavior in action for the high-complexity model in Fig. 1 — despite many large-variance regions, the next query is very close to the current best, and within our correlation bound. In Fig. 3b, we display the numerical solution to $\frac{\partial \mathtt{EI}}{\partial \rho} = 0$ together with the analytical bound in Proposition 4.1. We observe that for typical values of the GP mean and outputscale, the candidate query has substantial correlation with the incumbent under the aforementioned setting of an uninformed model.

With that said, the phenomenon of frequent querying of the boundary may still occur in practice, which we expand on in App. C.2. Specifically, querying of the boundary may occur when lengthscales are very long along one or more dimensions, which can occur when fitting the model using MLE (Williams & Rasmussen, 1995; Rasmussen & Williams, 2006; Karvonen & Oates, 2023). Then, candidates will be low-variance and highly correlated with existing data, despite being located at the boundary of one or multiple dimensions that are all deemed irrelevant. In Fig. 10 visualizes this setting for a 2D toy example.

Proposition 4.1 establishes that EI does not intrinsically pursue high-variance, uninformed regions. Rather, queries preferentially have substantial correlation with the incumbent. In a high-complexity setting, substantial correlation only arises when a data point is close to existing data, which suggests that EI should make queries in close proximity to the incumbent. As a result, the algorithm tends to very seldom query far from the incumbent, resulting in an exploitative behavior with similar qualities to local search. This phenomenon arises when there is negligible correlation in the data, namely when model complexity is too high to effectively model the objective function with existing data. In App. C.2, we expand on the local search behavior of the EI acquisition function, demonstrating its prevalence in conventional setups for dimensionalities as low as six.

4.3. Complexity of Existing HDBO

Having established that high complexity can yield uninformative models, as well as having discussed the link between complexity and dimensionality, it is evident that complexity assumptions must be sensible to facilitate a calibrated GP in HDBO. Notably, all the classes of HDBO algorithms outlined in Sec. 3 have such complexity-lowering assumptions. In Fig. 4, we display the modeled complexity of the most common classes of HDBO algorithms. Fig. 4 can be viewed as a cross section of Fig. 2, where we fix the number of data points to 1000 and instead vary the dimensionality of the problem to demonstrate how each HDBO class lower the growth of the complexity in the dimensionality, relative to a common global GP model as well as an independent kernel. The methods presented are: REMBO (Wang et al., 2016) with $d_e = 4$, random AddGPs (Ziomek & Bou Ammar, 2023), BOCK (Oh et al., 2018), Local GPs (Eriksson et al., 2019) after one round of shrinkage, the global GP with fixed lengthscales from Fig. 2, and our proposed method — scaling the lengthscales in the dimensionality of the problem, which is introduced later in Sec. 5. While each algorithm has parameters that affect the MIG, we have set parameters to make the comparison as fair as possible. In App. A.2, we outline the setup in detail.

As we have observed previously, employing a full-

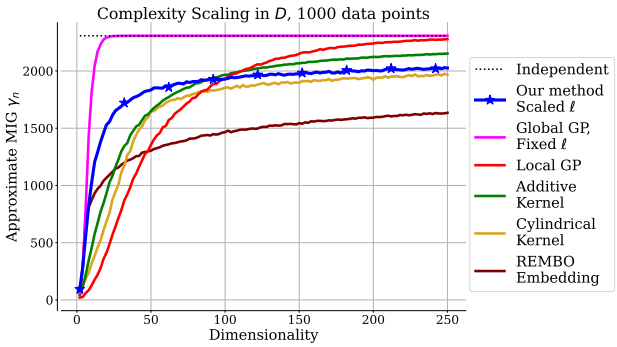


Figure 4. We display the model complexity scaling in the dimensionality of the problem for 1000 data points for various HDBO algorithms. Vanilla BO with fixed lengthscales (magenta) approaches independent complexity at approximately 20 dimensions. As expected, REMBO random embeddings (brown) reduce complexity the most, followed by BOCK cylindrical kernels (yellow). The MIG growth of our proposed modification of the global GP (blue) flattens out at a rate similar to cylindrical kernels (yellow), despite modelling the original, full-dimensional space.

dimensional GP without restrictions is far too complex, as we have approximate independence after 1000 observations already for an $18D$-objective. As expected, random subspace methods have small complexity increase in the ambient dimensionality. The increase stems from the fact that random, non-axis-aligned embeddings may slice the ambient dimensions very narrowly, which results in shorter lengthscales on the embedded model. The complexity increase for both our method (blue) and cylindrical kernels (yellow) stagnates rapidly, effectively assuming only marginal complexity increases after $D = 100$. Local methods (red) scale at same rates as global GPs (blue), but work on a drastically simplified model due to the lengthscale-scaled trust regions.

We re-iterate that low complexity is not strictly a desirable property, but as per Occam's razor, the most desirable property is to have the lowest possible complexity for a model that sufficiently aligns with the objective. This is especially true in the context of small data optimization, where each new data point acquired and employed to train the model is costly. We note, however, that the almost-independence exhibited by the global GP in Fig. 4 (magenta) for even moderate dimensionalities inevitably leads to the degeneracy highlighted in Sec. 4.2. Moreover, the assumptions behind each HDBO method are all means to the same end - reducing model complexity to manageable levels.

5. Low-complexity High-dimensional Bayesian Optimization

Hypothesizing that the pitfalls of HDBO are strictly caused by assumptions of insurmountable complexity, we present our main methodological contribution. We design a simple, plug-and-play assumption that retains almost constant complexity as the dimensionality increases. Similar to Fig. 1, we achieve this by adjusting the prior on the lengthscales to the dimensionality of the problem to the task at hand. Moreover, we ensure a calibrated signal variance by drawing on previous findings on GPs in an over-parameterized regime.

5.1. Ensuring Meaningful Correlation

Since stationary kernels compute covariances based on distances between data and both the diagonal and the distance between randomly sampled points in a D-dimensional hypercube scales as $\sqrt{D}$ (Köppen, 2000), increasing the lengthscales at this rate, $\ell_i \propto \sqrt{D}$, counteracts the complexity increase that stems from the increased distances. This change may, for example, be achieved by scaling the μ term of a LogNormal ($\mathcal{LN}$) prior

$$\ell_i \sim \mathcal{LN}\left(\mu_0 + \frac{\log(D)}{2}, \sigma_0\right) \tag{4}$$

where (μ_0, σ_0) are suitable parameters of $p(\boldsymbol{\ell})$ for a one-dimensional objective. This shifts both the mode and mean of the distribution by a factor of $\sqrt{D}$. Notably, our method does not increase the number of hyperparameters in a MAP-based BO setup. Furthermore, the proposed change may similarly be applied the more commonly used Gamma prior, with a different parameterization (Cho et al., 2004). Importantly, the change in complexity is not definitive, as we may still find some variables to be more important than others and adjust on-the-fly through MAP estimation of $\boldsymbol{\ell}$.

The proposed change suggests that the problem of large distances between points, and thereby the insurmountable complexity, is one that arises by assumption. Specifically, the lengthscale priors $p(\boldsymbol{\ell})$ employed by conventional BO frameworks (Balandat et al., 2020; Snoek et al., 2012b; Hutter et al., 2011) place substantial density on low values of $\boldsymbol{\ell}$. Assuming that all dimensions are of major importance may appear like a conservative and sensible choice. For moderately high dimensions, however, it practically guarantees that the problem will be impossible to model globally, even with the largest of budgets. Our method takes the opposite approach, and simply assumes that a problem *is simple enough to be modeled globally, for any dimensionality.*

5.2. Calibrating Epistemic Uncertainty

Lastly, we consider the role of the signal variance parameter, whose impact on data acquisition is evidenced by Prop. 4.1. Motivated by findings on on the optimal value $\hat{\sigma}_f^2$ of σ_f^2 generally in Moore et al. (2016) and in the over-parameterized regime by Ober et al. (2021), we consider

$$\hat{\sigma}_f^2 = \frac{1}{n}\boldsymbol{y}^{\mathrm{T}}\mathbf{K}^{-1}\boldsymbol{y} \tag{5}$$

which, in a BO context, has a different impact than in GPs generally. As we are able to selectively acquire our data, a large number of parameters and substantially correlated data will simplify data fit, driving down the optimal value of σ_f^2. When data is repeatedly re-normalized, the issue will be reinforced, as the signal variance is further decreased, and another highly correlated query is selected. As such, we fix $\sigma_f^2 = 1$ to match the scale of the standardized observations, and to ensure that σ_f^2 does not diminish over time. In App. F.1, we ablate this modeling choice and demonstrate its impact on the value of the outputscale during optimization for two real-world optimization tasks, and in App. F.2, we demonstrate the impact of the outputscale modeling choice on optimization performance.

6. Results

We now compare our Vanilla Bayesian Optimization method with a dimensionality-scaled lengthscale prior (we will refer to our method as D-scaled $p(\boldsymbol{\ell})$ or DSP for clarity), against state-of-the-art HDBO methods. We will include various classes of HDBO methods, such as the subspace-methods Bounce and SAASBO (Papenmeier et al., 2023; Eriksson & Jankowiak, 2021), the Local BO algorithms TuRBO (Eriksson et al., 2019) and Maximal Probability of Descent (Nguyen et al., 2022) (MPD), the AddGP method RD-UCB (Ziomek & Bou Ammar, 2023), the variable selection method MCTS-VS (Song et al., 2022), and CMA-ES (Hansen, 2006). We use each method's official repository, with the exception of SAASBO which is run through Ax (Bakshy et al.). We detail the experimental setup in App. A, and our code is publicly available at `https://github.com/hvarfner/vanilla_bo_in_highdim`.

We instantiate the DSP with $\mu_0 = \sqrt{2}, \sigma_0 = \sqrt{3}$, which equates to $\boldsymbol{\ell} \approx 0.50$ for $D = 6$ under the mode of $p(\boldsymbol{\ell})$. We initialize all methods with 30 samples, marked by a dashed vertical line. Bounce, CMA-ES and MPD deviate from conventional initialization. On all benchmarks, we use `LogEI` (Ament et al., 2023), using a low acquisition optimization budget of 512 initial (global) SOBOL samples and 512 Gaussian samples around the incumbent, followed by L-BFGS on the 4 best candidates, which is made possible by the low-complexity-high-smoothness model. In App. E, we demonstrate that our method delivers consistent performance for a range of complexities.

6.1. Sparse Synthetic Test Functions

We start by evaluating the DSP on a collection of commonly considered synthetic test functions with varying *total* and *effective dimensionality*. We note that the assumptions made in Sec. 5 diametrically oppose these test cases - each function has a low number of highly important dimensions with the large remainder being unimportant, whereas we assume that *each* dimension has relatively small impact.

The DSP is highly performant, as it rapidly identifies the important dimensions and subsequently optimizes the task. This is similar in to (Eriksson & Jankowiak, 2021), whose assumptions, represented through its sparse lengthscale prior, aligns perfectly with the task at hand. As such, SAASBO should, and does, perform best on average, with Vanilla BO being second. Notably, as Vanilla BO does not require the HMC (Bingham et al., 2018) fully Bayesian model fitting that SAASBO uses, it runs in a small fraction of the time.

6.2. A Plug-in on Mid-Dimensional Tasks

Subsequently, we use the DSP as a plug-in for low- and mid-dimensional tasks, primarily those considered in (Eriksson et al., 2019), to evaluate its ability to serve as a substitute for conventional, non-adaptive hyperparameter priors. The Lunar Lander (12D) and Robot Pushing (14D) tasks from (Wang et al., 2017), as well as the Swimmer (16D) and Hopper (32D) reinforcement learning tasks from the MuJoCo suite (Todorov et al., 2012), where we aim to learn a linear policy for two objects with varying degrees of freedom. We evaluate against a $\Gamma(3, 6)$ lengthscale prior with learnable σ_f^2, and against TuRBO, commonly considered the state-of-the-art mid-dimensional BO method. In Fig. 6, it is shown that the DSP is either competitive with, or outperforms, TuRBO on all tasks. In App. C, we empirically display the pronounced local search-like behavior of the $\Gamma(3, 6)$ prior for Hartmann (6D), which suggests that it does not offer a calibrated exploration-exploitation trade-off. On the contrary, the DSP maintains a moderate distance between queries, which indicates a calibrated trade-off throughout optimization.

6.3. High-dimensional Optimization Tasks

We now benchmark Vanilla BO with the DSP against a collection of frequently considered tasks in the high-dimensional literature (Eriksson & Jankowiak, 2021; Eriksson et al., 2019; Papenmeier et al., 2022; 2023; Šehić et al., 2021): Specifically, we consider MOPTA08 (124D), SVM (388D), Lasso-DNA (180D), and the MuJoCo (Todorov et al., 2012) Ant (888D) and Humanoid (6392D) reinforcement learning tasks. We stress that, for all benchmarks where applicable, (BAxUS on SVM, Lasso-DNA and Humanoid, TuRBO on MOPTA and SVM, RD-UCB on Lasso-DNA), baselines perform within the error bars of the original implementation (Papenmeier et al., 2022; Ziomek & Bou Ammar, 2023) or in other papers by the same authors in the case of TuRBO (Eriksson et al., 2019; Eriksson & Jankowiak, 2021). In App. D, we display the performance in a lower-budget setting. In App. E.1, we compare to conventional MAP and MLE. Fig. 7 shows that Vanilla BO

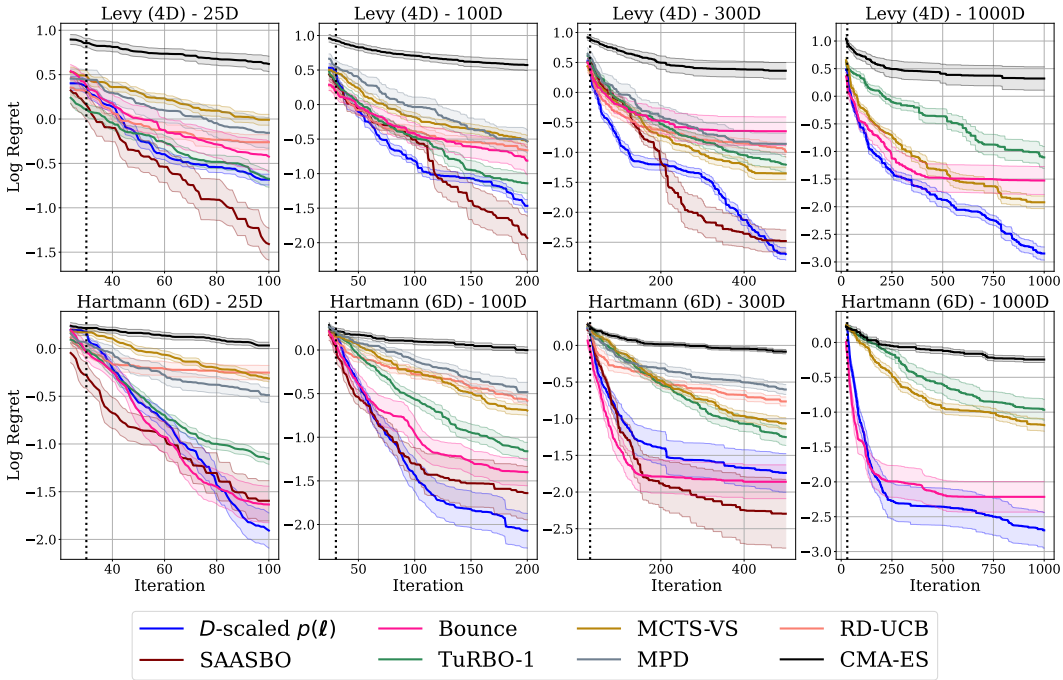


Figure 5. Average log regret of all baselines on Levy (4D) and Hartmann (6D) synthetic test functions of varying ambient dimensionality across 20 repetitions (10 for SAASBO). Vanilla BO performs second best, beaten only by SAASBO on four tasks, whose axis-aligned subspace assumption (along with MCTS-VS' variable selection) aligns perfectly with the task at hand. We omit SAASBO from the 1000D benchmarks due to the prohibitive runtime, and RD-UCB and MPD due to a combination of runtime and numerical instability.

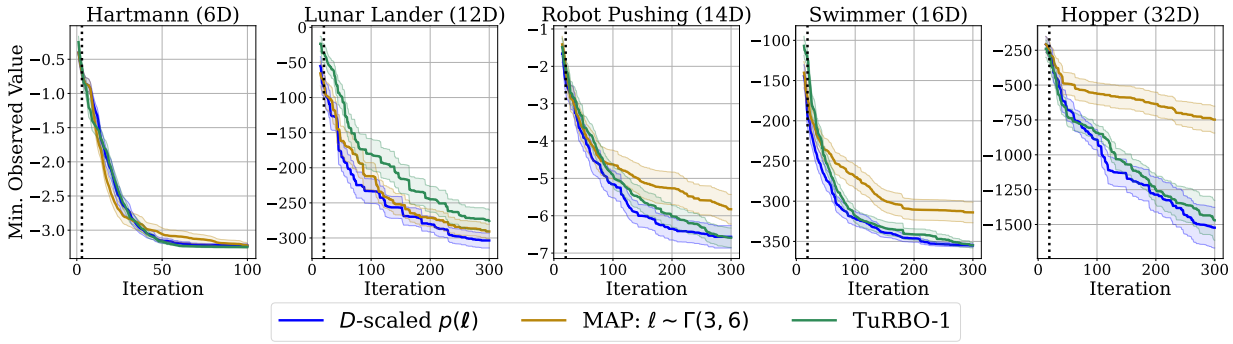


Figure 6. Best observed value for the DSP, conventional MAP, and TuRBO on Hartmann (6D) and four mid-dimensional real-world optimization tasks. All methods perform comparably on Hartmann, while the DSP outperforms or is on part with TuRBO on the other tasks. Notably, the DSP performs at least equally well as the $\Gamma(3, 6)$ on all tasks, and substantially better on 4 out of 5 tasks, which suggests that it is well-suited as a drop-in replacement for conventional priors.

with the DSP is highly competitive, performing the best by a substantial margin on Lasso-DNA and Humanoid, and produces top-two performance on the remaining tasks. On the MuJuCo Ant, Bounce's low-dimensional initialization allows it to obtain an average value of 800 after DoE due to consistently sampling data points close to the center of the search space. Notably, the DSP is very consistent between repetitions, as evident by the small error bars. This can be attributed to the consistent modelling, as the DSP is not dependent on randomness in subspace design, trust region initialization or variable selection. Rather, it obtains a consistent, calibrated model through meaningful correlation in the data, from which it can effectively infer promising regions and improve upon the DoE.

Notably, the DSP does *not* heavily depend on identification of active variables. In Fig. 8, we demonstrate for the distribution of lengthscale values for the 388D SVM after 250 iterations. The DSP does not consistently identify active dimensions with large confidence. Instead, the calibrated complexity of the model allows for meaningful inference

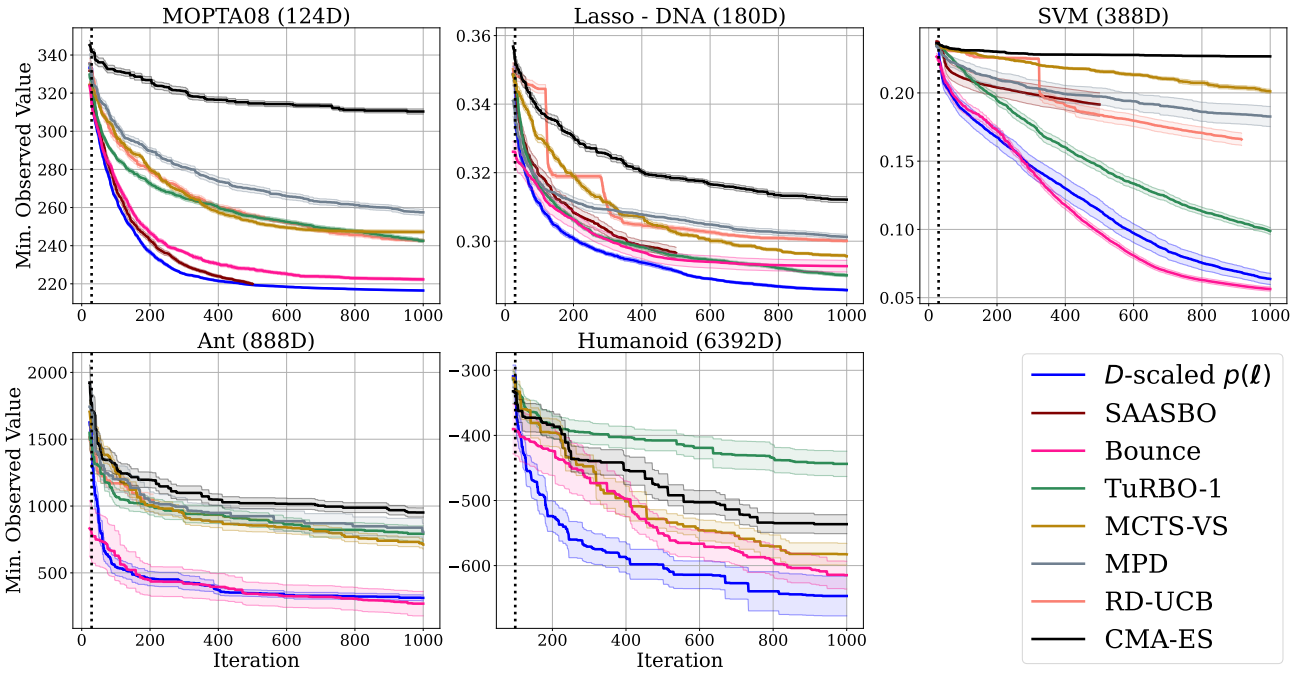


Figure 7. Best observed value of all baselines on five real-world tasks from of various domains across 20 repetitions (10 for SAASBO). Vanilla BO performs best across all tasks except for MOPTA, where it is outperformed by SAASBO, and the MuJoCo Ant, where BAxUS gets an advantage from performing the initialization phase in a lower-dimensional subspace, which enables it to obtain initial samples close to the center of the search space. Vanilla BO substantially outperforms all baselines on Lasso-DNA, SVM and Humanoid. Notably, the extreme dimensionality of Humanoid does not have an apparent negative impact the performance of Vanilla BO.

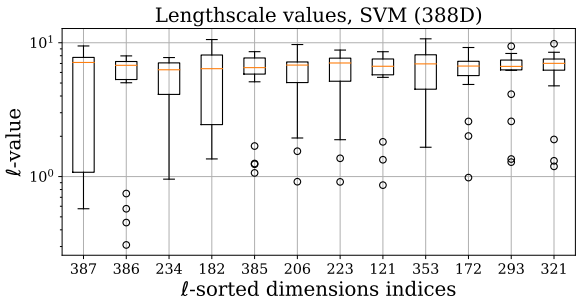


Figure 8. Distribution of lengthscale values for the DSP on the 388D SVM task. Lengthscales are sorted according to their mean log value. The three last indexed dimensions (385, 386, 387) are considered particularly active (Eriksson & Jankowiak, 2021), but are not identified as such consistently in our method. the black horizontal lines indicate upper and lower quartiles, and the orange horizontal lines indicate the median.

along all dimensions, until particularly active dimensions are potentially identified. As such, the DSP does not *require* the identification of active variables to achieve calibrated BO, but its identification helps optimization. As such, we attribute the superior performance of our method to the calibrated complexity, and the effective inference and exploration-exploitation trade-off that stems from it.

7. Conclusion and Future Work

The curse of dimensionality has long been assumed to hinder the application of conventional Bayesian optimization in high dimensions. We show that the hindrance is not driven by dimensionality, but by the assumed complexity of the objective. We make minor modifications to the assumptions of the vanilla BO algorithm to make complexity scaling manageable with increasing dimensionality. As a result, we demonstrate that vanilla BO is extremely effective for problems of high dimensionality, outperforming the state-of-the-art for problems with dimensions into the thousands.

Nevertheless, we do not believe that tailored high-dimensional BO algorithms are unwarranted: if the problem at hand is known to adhere to the structural assumptions that are conventionally made (effective subspace, additivity) or where non-stationarity in the objective facilitates local modelling, we believe these approaches will be superior to the vanilla algorithm. However, these restrictive assumptions should not be made *out of necessity*, but when prior knowledge supports them. For future work, we plan to investigate the topic of complexity as it relates to modelling in Bayesian optimization more broadly, and in the context of latent space GP models (Griffiths & Hernández-Lobato, 2017; Maus et al., 2022).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

Carl Hvarfner, Erik Hellsten and Luigi Nardi were partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Luigi Nardi was partially supported by the Wallenberg Launch Pad (WALP) grant nbr. 2021.0348.

References

Ament, S., Daulton, S., Eriksson, D., Balandat, M., and Bakshy, E. Unexpected improvements to expected improvement for bayesian optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1vyAG6j9PE.

Bakshy, E., Dworkin, L., Karrer, B., Kashin, K., Letham, B., Murthy, A., and Singh, S. Ae: A domain-agnostic platform for adaptive experimentation.

Balandat, M., Karrer, B., Jiang, D. R., Daulton, S., Letham, B., Wilson, A. G., and Bakshy, E. Botorch: A framework for efficient monte-carlo bayesian optimization. In *Advances in Neural Information Processing Systems*, 2020. URL http://arxiv.org/abs/1910.06403.

Baptista, R. and Poloczek, M. Bayesian optimization of combinatorial structures. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 462–471. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/baptista18a.html.

Berkenkamp, F., Schoellig, A. P., and Krause, A. No-regret bayesian optimization with unknown hyperparameters. *Journal of Machine Learning Research*, 20(50):1–24, 2019. URL http://jmlr.org/papers/v20/18-213.html.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 2018.

Binois, M. and Wycoff, N. A survey on high-dimensional gaussian process modeling with application to bayesian optimization. *ACM Trans. Evol. Learn. Optim.*, 2(2), aug 2022. doi: 10.1145/3545611. URL https://doi.org/10.1145/3545611.

Binois, M., Ginsbourger, D., and Roustant, O. On the choice of the low-dimensional domain for global optimization via random embeddings. *J. of Global Optimization*, 76 (1):69–90, jan 2020. ISSN 0925-5001. doi: 10.1007/s10898-019-00839-1. URL https://doi.org/10.1007/s10898-019-00839-1.

Bull, A. D. Convergence rates of efficient global optimization algorithms. 12:2879–2904, 2011.

Cho, H.-K., Bowman, K. P., and North, G. R. A Comparison of Gamma and Lognormal Distributions for Characterizing Satellite Rain Rates from the Tropical Rainfall Measuring Mission. *Journal of Applied Meteorology*, 43(11): 1586–1597, November 2004. doi: 10.1175/JAM2165.1.

De Ath, G., Fieldsend, J. E., and Everson, R. M. What do you mean? the role of the mean function in bayesian optimisation. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion*, GECCO '20, pp. 1623–1631, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371278. doi: 10.1145/3377929.3398118. URL https://doi.org/10.1145/3377929.3398118.

Djolonga, J., Krause, A., and Cevher, V. High-dimensional gaussian process bandits. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/8d34201a5b85900908db6cae92723617-Paper.pdf.

Duembgen, L. Bounding standard gaussian tail probabilities. *arXiv: Statistics Theory*, 2010. URL https://api.semanticscholar.org/CorpusID:88512442.

Duvenaud, D. K., Nickisch, H., and Rasmussen, C. Additive gaussian processes. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/4c5bde74a8f110656874902f07378009-Paper.pdf.

Eduardo, A. and Gutmann, M. U. Bayesian optimization with informative covariance. *Trans. Mach. Learn. Res.*, 2023, 2022. URL https://api.semanticscholar.org/CorpusID:251320397.

Eriksson, D. and Jankowiak, M. High-dimensional Bayesian optimization with sparse axis-aligned subspaces. In de Campos, C. and Maathuis, M. H. (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 493–503. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/eriksson21a.html.

Eriksson, D., Pearce, M., Gardner, J., Turner, R. D., and Poloczek, M. Scalable global optimization via local bayesian optimization. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/6c990b7aca7bc7058f5e98ea909e924b-Paper.pdf.

Gardner, J., Guo, C., Weinberger, K., Garnett, R., and Grosse, R. Discovering and Exploiting Additive Structure for Bayesian Optimization. In Singh, A. and Zhu, J. (eds.), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pp. 1311–1319. Proceedings of Machine Learning Research, 2017.

Garnett, R., Osborne, M. A., and Hennig, P. Active learning of linear embeddings for gaussian processes. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pp. 230–239, Arlington, Virginia, USA, 2014. AUAI Press. ISBN 9780974903910.

GPyOpt-authors, T. GPyOpt: A bayesian optimization framework in python. http://github.com/SheffieldML/GPyOpt, 2016.

Griffiths, R.-R. and Hernández-Lobato, J. M. Constrained bayesian optimization for automatic chemical design. *arXiv: Machine Learning*, 2017.

Han, E., Arora, I., and Scarlett, J. High-dimensional bayesian optimization via tree-structured additive models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7630–7638, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16933.

Hansen, N. The CMA evolution strategy: a comparing review. In Lozano, J., Larranaga, P., Inza, I., and Bengoetxea, E. (eds.), *Towards a new evolutionary computation. Advances on estimation of distribution algorithms*, pp. 75–102. 2006.

Hellsten, E. O., Hvarfner, C., Papenmeier, L., and Nardi, L. High-dimensional bayesian optimization with group testing, 2023.

Hutter, F., Hoos, H. H., and Leyton-Brown, K. Sequential model-based optimization for general algorithm configuration. In *Learning and Intelligent Optimization*, 2011.

Hvarfner, C., Hellsten, E. O., Hutter, F., and Nardi, L. Self-correcting bayesian optimization through bayesian active learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=dX9MjUtP1A.

Jones, D., Schonlau, M., and Welch, W. Efficient global optimization of expensive black box functions. 13:455–492, 1998.

Jones, D. R. A taxonomy of global optimization methods based on response surfaces. 21:345–383, 2001.

Kandasamy, K., Schneider, J., and Póczos, B. High Dimensional Bayesian Optimisation and Bandits via Additive Models. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning (ICML'15)*, volume 37, pp. 295–304. Omnipress, 2015.

Karvonen, T. and Oates, C. J. Maximum likelihood estimation in gaussian process regression is ill-posed. *Journal of Machine Learning Research*, 24(120):1–47, 2023. URL http://jmlr.org/papers/v24/22-1153.html.

Kirschner, J., Mutny, M., Hiller, N., Ischebeck, R., and Krause, A. Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3429–3438. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kirschner19a.html.

Köppen, M. The curse of dimensionality. *5th online world conference on soft computing in industrial applications (WSC5)*, 2000.

Krause, A., Singh, A., and Guestrin, C. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(8):235–284, 2008. URL http://jmlr.org/papers/v9/krause08a.html.

Letham, B., Calandra, R., Rai, A., and Bakshy, E. Re-examining linear embeddings for high-dimensional Bayesian optimization. *Advances in neural information processing systems*, 33:1546–1558, 2020.

Li, C., Gupta, S., Rana, S., Nguyen, V., Venkatesh, S., and Shilton, A. High dimensional bayesian optimization using dropout. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pp. 2096–2102. AAAI Press, 2017. ISBN 9780999241103.

Malu, M., Dasarathy, G., and Spanias, A. Bayesian optimization in high-dimensional spaces: A brief survey. In *2021 12th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–8, 2021. doi: 10.1109/IISA52424.2021.9555522.

Maus, N., Jones, H., Moore, J., Kusner, M. J., Bradshaw, J., and Gardner, J. Local latent space bayesian optimization over structured inputs. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 34505–34518. Curran Associates, Inc., 2022.

Moore, C. J., Chua, A. J. K., Berry, C. P. L., and Gair, J. R. Fast methods for training gaussian processes on large datasets. *Royal Society Open Science*, 3(5):160125, May 2016. ISSN 2054-5703. doi: 10.1098/rsos.160125. URL http://dx.doi.org/10.1098/rsos.160125.

Müller, S., von Rohr, A., and Trimpe, S. Local policy search with bayesian optimization. In *Advances in Neural Information Processing Systems*, 2021.

Nayebi, A., Munteanu, A., and Poloczek, M. A framework for Bayesian optimization in embedded subspaces. In *International Conference on Machine Learning*, pp. 4752–4761. PMLR, 2019.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. An analysis of approximations for maximizing submodular set functions–i. *Math. Program.*, 14(1):265–294, dec 1978. ISSN 0025-5610. doi: 10.1007/BF01588971. URL https://doi.org/10.1007/BF01588971.

Nguyen, Q., Wu, K., Gardner, J., and Garnett, R. Local bayesian optimization via maximizing probability of descent. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 13190–13202. Curran Associates, Inc., 2022.

Ober, S. W., Rasmussen, C. E., and van der Wilk, M. The promises and pitfalls of deep kernel learning. In de Campos, C. and Maathuis, M. H. (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1206–1216. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/ober21a.html.

Oh, C., Gavves, E., and Welling, M. BOCK : Bayesian optimization with cylindrical kernels. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3868–3877. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/oh18a.html.

Osborne, M. A. *Bayesian Gaussian processes for sequential prediction, optimisation and quadrature*. PhD thesis, Oxford University, UK, 2010.

Papenmeier, L., Nardi, L., and Poloczek, M. Increasing the scope as you learn: Adaptive bayesian optimization in nested subspaces. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=e4Wf6112DI.

Papenmeier, L., Nardi, L., and Poloczek, M. Bounce: a Reliable Bayesian Optimization Algorithm for Combinatorial and Mixed Spaces. *arXiv preprint arXiv:2307.00618*, 2023.

Rana, S., Li, C., Gupta, S., Nguyen, V., and Venkatesh, S. High dimensional Bayesian optimization with elastic Gaussian process. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2883–2891. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/rana17a.html.

Rasmussen, C. and Williams, C. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Rolland, P., Scarlett, J., Bogunovic, I., and Cevher, V. High-dimensional bayesian optimization via additive models with overlapping groups. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 298–307. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/rolland18a.html.

Rothfuss, J., Fortuin, V., Josifoski, M., and Krause, A. Pacoh: Bayes-optimal meta-learning with pac-guarantees. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9116–9126. PMLR, 18–24 Jul 2021a. URL https://proceedings.mlr.press/v139/rothfuss21a.html.

Rothfuss, J., Heyn, D., Chen, J., and Krause, A. Meta-learning reliable priors in the function space. In *Advances in Neural Information Processing Systems*, volume 34, 2021b.

Šehić, K., Gramfort, A., Salmon, J., and Nardi, L. LassoBench: A High-Dimensional Hyperparameter Optimization Benchmark Suite for Lasso. *arXiv preprint arXiv:2111.02790*, 2021.

Siivola, E., Vehtari, A., Vanhatalo, J. P., González, J. I., and Andersen, M. R. Correcting boundary over-exploration deficiencies in bayesian optimization with virtual derivative sign observations. *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2017. URL https://api.semanticscholar.org/CorpusID:53236252.

Snoek, J., Larochelle, H., and Adams, R. Practical Bayesian optimization of machine learning algorithms. In Bartlett, P., Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems (NeurIPS'12)*, pp. 2960–2968, 2012a.

Snoek, J., Larochelle, H., and Adams, R. P. Practical bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems, pp. 2951–2959, Red Hook, NY, USA, 2012b. Curran Associates Inc.

Snoek, J., Swersky, K., Zemel, R., and Adams, R. Input warping for Bayesian optimization of non-stationary functions. In Xing, E. and Jebara, T. (eds.), *Proceedings of the 31th International Conference on Machine Learning, (ICML'14)*, pp. 1674–1682. Omnipress, 2014.

Song, L., Xue, K., Huang, X., and Qian, C. Monte carlo tree search based variable selection for high dimensional bayesian optimization. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=SUzPos_pUC.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. In Fürnkranz, J. and Joachims, T. (eds.), *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*, pp. 1015–1022. Omnipress, 2010.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, May 2012. ISSN 1557-9654. doi: 10.1109/tit.2011.2182033. URL http://dx.doi.org/10.1109/TIT.2011.2182033.

Swersky, K., Duvenaud, D., Snoek, J., Hutter, F., and Osborne, M. A. Raiders of the lost architecture: Kernels for bayesian optimization in conditional parameter spaces, 2014.

Swersky, K. J. *Improving Bayesian Optimization for Machine Learning using Expert Priors*. PhD thesis, 2017.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Wan, X., Nguyen, V., Ha, H., Ru, B., Lu, C., and Osborne, M. A. Think global and act local: Bayesian optimisation over high-dimensional categorical and mixed search spaces. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10663–10674. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wan21b.html.

Wang, L., Fonseca, R., and Tian, Y. Learning search space partition for black-box optimization using monte carlo tree search. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19511–19522. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e2ce14e81dba66dbff9cbc35ecfdb704-Paper.pdf.

Wang, Z., Hutter, F., Zoghi, M., Matheson, D., and De Feitas, N. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55:361–387, 2016.

Wang, Z., Li, C., Jegelka, S., and Kohli, P. Batched high-dimensional Bayesian optimization via structural kernel learning. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3656–3664. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/wang17h.html.

Williams, C. and Rasmussen, C. Gaussian processes for regression. In Touretzky, D., Mozer, M., and Hasselmo, M. (eds.), *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995. URL https://proceedings.neurips.cc/paper_files/paper/1995/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

Wu, K., Kim, K., Garnett, R., and Gardner, J. R. The behavior and convergence of local bayesian optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=9KtX12YmA7.

Yao, Y., Vehtari, A., and Gelman, A. Stacking for non-mixing bayesian computations: The curse and blessing of multimodal posteriors, 2020.

Ziomek, J. K. and Bou Ammar, H. Are random decompositions all we need in high dimensional Bayesian optimisation? In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43347–43368. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ziomek23a.html.

A. Experimental Setup

We outline our experimental setup, detailing the specifics of the baseline methods. Our code is publically available at `https://github.com/hvarfner/vanilla_bo_in_highdim`.

A.1. SAASBO

Due to the computational expense of SAASBO model fitting and subsequent forward passes through the acquisition function, we run the lower-budget variant outlined in App. A1 of (Eriksson & Jankowiak, 2021). Due to the 16-fold increase in the cost of the forward pass in the fully Bayesian setting, we limit the maximal number of evaluations to 500. Lastly, the model is retrained once every 4 iterations, since a single model fitting takes upwards of 5 minutes on SVM on 4 CPUS for later iterations.

A.2. Complexity Comparison Setup

For the complexity comparison in Fig. 4, we benchmark against the following instantiations of each HDBO algorithm:

Cylindrical Kernel We stay as close as possible to the default values presented in BOCK (Oh et al., 2018), using a degree-3 polynomial for the angular component of the kernel. We disregard input warpings (Snoek et al., 2014), as we do not view them as related to the dimensionality-reducing component of the model.

AddGP We randomly sample groups by sequentially considering one dimension at a time. With equal probability for each existing group and one empty group, the dimension under consideration is added to one of the groups. This setup is similar, but not equivalent, to the algorithm by Ziomek & Bou Ammar (2023).

Local GP We consider the instantiation described by Eriksson et al. (2019) of constructing the trust region with a size of $0.8\boldsymbol{\ell}$. Since we plot the complexity of the local GP after one round of shrinkage, the trust region side length is equal to $0.4\boldsymbol{\ell}$. Moreover, we assume this trust region to hold *all* data to enable a comparison with other methods.

Random Embedding We instantiate the standard REMBO (Wang et al., 2016) embedding from a D-dimensional Gaussian distribution. We compute the complexity by based on the shortest lengthscale of each subspace dimension. These, in turn, are computed from by multiplying the base lengthscale, $\ell = 0.5$, by the fraction of the effective dimension that the ambient dimension is occupying, as visualized in Fig.3 of Wang et al. (2016).

B. Benchmarks

We outline the exact benchmarks used, including the exact bounds and the noise level. For Levy, we offset each active dimension. Since the optimum is in $[1, 1, 1, \ldots]$, methods that search along diagonals in the search space, such as BAxUS (Papenmeier et al., 2022), gets an outsized, and arguably unrealistic, advantage.

Notably, We use the BAxUS variant of SVM, which uses a smaller subset than the original benchmark by (Eriksson & Jankowiak, 2021) for faster evaluation. As evident form a comparison of the results in Papenmeier et al. (2022) and (Eriksson & Jankowiak, 2021), these benchmarks produce different results, suggesting that the objective functions are substantially different.

Task	Effective Dimensionality	σ_ϵ	Search space
Levy (4D)	4	0.01	$[-10, 5] \times [-10, 10] \times [-5, 10] \times [-1, 10]$
Hartmann (6D)	6	0.01	$[0, 1]^D$

Table 1. Benchmarks used for the synthetic, axis-aligned experiments.

C. The Locality Issue

We outline our findings related to the boundary and locality issue, proving that the original description (Swersky, 2017) does not happen in practice. Subsequently, we display our empirical findings regarding the locality issue for BO with a

Task	Dimensionality	Search space
MOPTA	124	$[0,1]^D$
Lasso-DNA	180	$[-1,1]^D$
SVM	388	$[0,1]^D$
Ant	888	$[-1,1]^D$
Humanoid	6392	$[-1,1]^D$

Table 2. Benchmarks used for the real-world optimization experiments.

conventional, short-lengthscale hyperprior using the BoTorch (Balandat et al., 2020) default. We note, though, that other BO frameworks (Hutter et al., 2011; Snoek et al., 2012b; GPyOpt-authors, 2016) employ lengthscale priors with very similar characteristics, and any of them could have been used for our purpose.

C.1. Theory: Non-existence of the Boundary Issue

To prove that EI will never prefer a point with maximal variance under the proposed setup, we first restate the formal definition of EI:

$$\texttt{EI}(\boldsymbol{x}) = \mathbb{E}\left[[(f(\boldsymbol{x}) - y_{\max}]^+\right] = Z\sigma(\boldsymbol{x})\Phi(Z) + \sigma(\boldsymbol{x})\phi(Z), \tag{6}$$

where y_{max} is the maximal objective function value observed so far, $Z = (\mu(\boldsymbol{x}) - y_{\max})/\sigma(\boldsymbol{x})$, and ϕ and Φ are the PDF and CDF of the standard Gaussian, respectively. Hence, for a noiseless function the next point $\boldsymbol{x}_*$ to query is

$$\boldsymbol{x}_* \in \arg\max_{\boldsymbol{x}\in\mathcal{X}} \texttt{EI}(\boldsymbol{x}). \tag{7}$$

Note that we are considering a maximization problem. Lastly, we denote the mean of the GP by c.

Proposition C.1. *Assume that $y_{max} > c$, $\mathbf{K} = \boldsymbol{I}$ and that the candidate query $\boldsymbol{x}_*$ correlates with at most one observation. Then, the correlation $\rho^* = \sigma_f^{-2} k(\boldsymbol{x}_*, \boldsymbol{x}_{inc})$ between the next query $\boldsymbol{x}_* = \arg\max_{\boldsymbol{x}\in\mathcal{X}} EI(\boldsymbol{x})$ and $\boldsymbol{x}_{inc}$ satisfies*

$$\rho^* \sqrt{\frac{1+\rho^*}{1-\rho^*}} \geqslant \frac{y_{max} - c}{\sigma_f}. \tag{8}$$

Proof. Since $\mathbf{K} = \sigma_f^2 \boldsymbol{I}$ and $\boldsymbol{x}_*$ correlates with only one observation, we need only consider the proximity of $\boldsymbol{x}_*$ to the incumbent $\boldsymbol{x}_{inc}$. As such, we start by parametrizing EI by the correlation ρ , so that $k(\boldsymbol{x}_*, \boldsymbol{x}_{inc}) = \sigma_f^2 \rho$:

$$\mu(\boldsymbol{x}_*) = c + k(\boldsymbol{x}_*, \boldsymbol{X}) k(\boldsymbol{X}, \boldsymbol{X})^{-1} (\boldsymbol{y} - c) = c + \rho(y_{\max} - c)$$

$$\sigma^2(\boldsymbol{x}_*) = k(\boldsymbol{x}_*, \boldsymbol{x}_*) + k(\boldsymbol{x}_*, \boldsymbol{X}) k(\boldsymbol{X}, \boldsymbol{X})^{-1} k(\boldsymbol{X}, \boldsymbol{x}_*) = \sigma_f^2 (1 - \rho^2). \tag{9}$$

We will by $\hat{\texttt{EI}}(\rho)$ denote the ρ-parametrized EI, which is a function over ρ, to distinguish it from the regular $\texttt{EI}(\boldsymbol{x})$, which is a function over $\boldsymbol{x}$. Moreover, for ease of notation, we scale our data such that $c = 0$, $\sigma_f^2 = 1$, and $\hat{y}_{\max} = \frac{y_{\max} - c}{\sigma_f} > 0$. After scaling we further get $\hat{Z} = \left(\frac{(\rho-1)}{\sqrt{1-\rho^2}} \hat{y}_{\max}\right)$, and finally:

$$\hat{\texttt{EI}}(\rho) = (\rho - 1)\, \hat{y}_{\max} \Phi\left(\frac{(\rho-1)}{\sqrt{1-\rho^2}} \hat{y}_{\max}\right) + \sqrt{1-\rho^2}\, \phi\left(\frac{(\rho-1)}{\sqrt{1-\rho^2}} \hat{y}_{\max}\right). \tag{10}$$

Differentiating with regard to ρ, we obtain that:

$$\frac{\partial \hat{\texttt{EI}}}{\partial \rho} = \hat{y}_{\max} \Phi(\hat{Z}) - \frac{\rho}{\sqrt{1-\rho^2}} \phi(\hat{Z}). \tag{11}$$

We then establish the values of ρ for which $\frac{\partial \mathtt{EI}}{\partial \rho}$ are positive. We bound $\Phi(\hat{Z})$ as a function of $\phi(\hat{Z})$ using known inequalities for Mills' ratio $(1-\Phi)/\phi$ (Duembgen, 2010),

$$\frac{2\phi(\hat{Z})}{\sqrt{4+\hat{Z}^2}-\hat{Z}} \leqslant \Phi(\hat{Z}), \quad -\hat{y}_{\text{max}} \leqslant \hat{Z} \leqslant 0.$$

We then insert the inequality into the expression for $\frac{\partial \hat{\mathtt{EI}}}{\partial \rho}$, which leaves

$$\begin{aligned}
\frac{\partial \hat{\mathtt{EI}}}{\partial \rho} &\geqslant \left[\frac{2\hat{y}_{\text{max}}}{\sqrt{4+\hat{Z}^2}-\hat{Z}} - \frac{\rho}{\sqrt{1-\rho^2}}\right]\phi(\hat{Z}) \\
&= \left[\frac{2\hat{y}_{\text{max}}}{\sqrt{4+\frac{(1-\rho)^2}{1-\rho^2}\hat{y}_{\text{max}}^2}+\frac{1-\rho}{\sqrt{1-\rho^2}}\hat{y}_{\text{max}}} - \frac{\rho}{\sqrt{1-\rho^2}}\right]\phi(\hat{Z}) \\
&\geqslant \left[\frac{2\hat{y}_{\text{max}}}{2+2\hat{y}_{\text{max}}\frac{1-\rho}{\sqrt{1-\rho^2}}} - \frac{\rho}{\sqrt{1-\rho^2}}\right]\phi(\hat{Z}),
\end{aligned}$$

where we apply the triangle inequality to the denominator of the left term between the first and second step. We wish to lower bound $\rho^* = \arg\max \hat{\mathtt{EI}}(\rho)$ by finding the values of ρ for which $\hat{\mathtt{EI}}$ is strictly increasing. Thus, we set our lower bound for the derivative to zero and obtain

$$\frac{\partial \hat{\mathtt{EI}}}{\partial \rho} \geqslant \left[\frac{\hat{y}_{\text{max}}}{1+\hat{y}_{\text{max}}\frac{1-\rho}{\sqrt{1-\rho^2}}} - \frac{\rho}{\sqrt{1-\rho^2}}\right]\phi(\hat{Z}) \geqslant 0.$$

We then multiply by $\sqrt{1-\rho^2}$, and divide by $\phi(\hat{Z})$, which are both > 0 for $\rho > 0$, and re-arrange:

$$\begin{aligned}
&\frac{\hat{y}_{\text{max}}(1-p^2)}{\sqrt{1-\rho^2}+\hat{y}_{\text{max}}(1-\rho)} - \rho \geqslant 0 \iff \\
&\frac{\hat{y}_{\text{max}}(1+\rho)}{\sqrt{\frac{1+\rho}{1-\rho}}+\hat{y}_{\text{max}}} - \rho \geqslant 0 \iff \\
&\hat{y}_{\text{max}} - \rho\sqrt{\frac{1+\rho}{1-\rho}} \geqslant 0
\end{aligned}$$

As such, we can conclude, that $\hat{\mathtt{EI}}$ is strictly increasing in ρ as long as $\hat{y}_{\text{max}} \geqslant \rho\sqrt{\frac{1+\rho}{1-\rho}}$, meaning that $\rho^* = \arg\max_{\rho\in[0,1)} \hat{\mathtt{EI}}(\rho)$ satisfies the inequality

$$\rho^*\sqrt{\frac{1+\rho^*}{1-\rho^*}} \geqslant \hat{y}_{\text{max}} = \frac{y_{\text{max}}-c}{\sigma_f}.$$

□

C.2. The Locality Issue in Practice

We now demonstrate the practical prevalence of the locality issue, expanding on the claims from Sec. 4.2. We note that the locality issue applies to the high-complexity setting, as generally outlined in Sec. 4. While the high-complexity setting is most prevalent when modeling high-dimensional problems, we will exemplify the locality issue through low-dimensional examples where complexity has been artificially increased by shortening the lengthscale of the model.

In Fig. 9, we demonstrate the continuation of Fig. 3 for six iterations. We visualize the bound on ρ^* in green for the first three iterations (top row) when the incumbent is not in the interior of the good region of the search space. Notably,

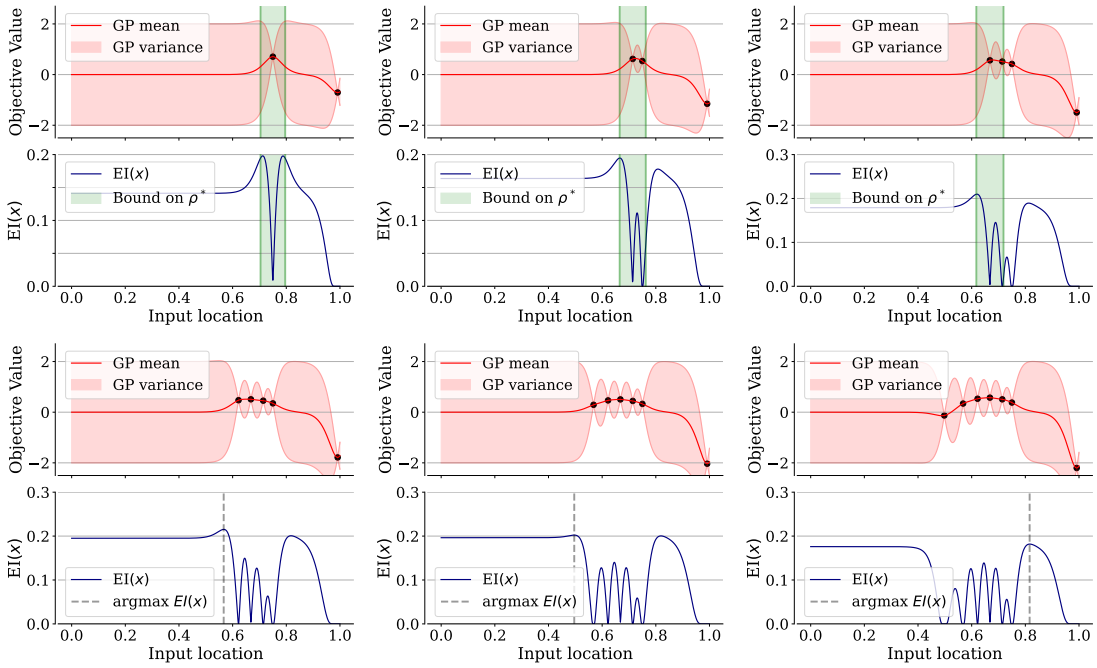


Figure 9. Visualization of the locality issue for the continuation of Fig. 3, considering the high-complexity model with short lengthscales. For the first three iterations (top row), we visualize the bound on ρ^* since the incumbent is at the leftmost edge of the well-performing region. In the bottom row, the local behavior continues despite no longer observing improvement. Notably, the large variance region in the leftmost part of the search space is never preferred by EI despite continuous re-normalization of the data.

the large-variance region in the leftmost part of the search space is never considered to be the best, despite repeated re-normalization of the data between iterations.

We re-iterate that the EI-value of the large-variance regions may often be *good, but not optimal*, which is clearly exemplified in Fig. 9.

C.3. The Boundary Issue through the Lens of Locality

We now demonstrate how the boundary-seeking behavior (Swersky, 2017) may occur even for a complex model. Importantly, this occurs under special circumstances, when a small number of dimensions account for a large degree of the total model complexity, with other dimensions being practically deactivated. In a high-dimensional setting, where the number of GP lengthscales may be on the same order of magnitude as the number of data points, this becomes increasingly likely.

In Fig. 10, we visualize how EI may choose to query high-correlation points that are seemingly distant, and even along the boundary of the search space. We visualize three observations (yellow), the posterior moments, and the value of EI across the search space. The model has a short lengthscale in x_2, but the large lengthscale along the x_1 dimension allows for traversal across the dimension with very little change in predictions and correlation. For this setting, the optimal query (green star) is close to the incumbent, within our correlation bound. However, all of the green line constitute queries that are within a factor 10^-5 of the $\max$ of EI. Queries along this line are decidedly *not* exploratory ones (in the classical high-variance sense), since they have very high correlation with the incumbent. However, they cover both boundaries, and a large region outside the bounds of the search space. Due to the instability of MLE in the overparametrized, high-dimensional setting, long-lengthscale may appear frequently dimensions, and can also vary substantially between model fittings.

Lastly, we demonstrate the drastically different behaviors that may arise even in low-dimensional problems, such as the non-embedded variant of Hartmann (6D). Since our enhancement is applicable to low-dimensional problems as well, we demonstrate its ability to exhibit a well-calibrated exploration-exploitation trade-off. We demonstrate the performance of MLE, conventional (Balandat et al., 2020), low-lengthscale MAP ($\ell_i \sim \Gamma(3, 6)$) and the DSP in Fig. 11 While conventional MAP and DSP achieve almost identical final performance, we now consider the degree to which the locality issue is present

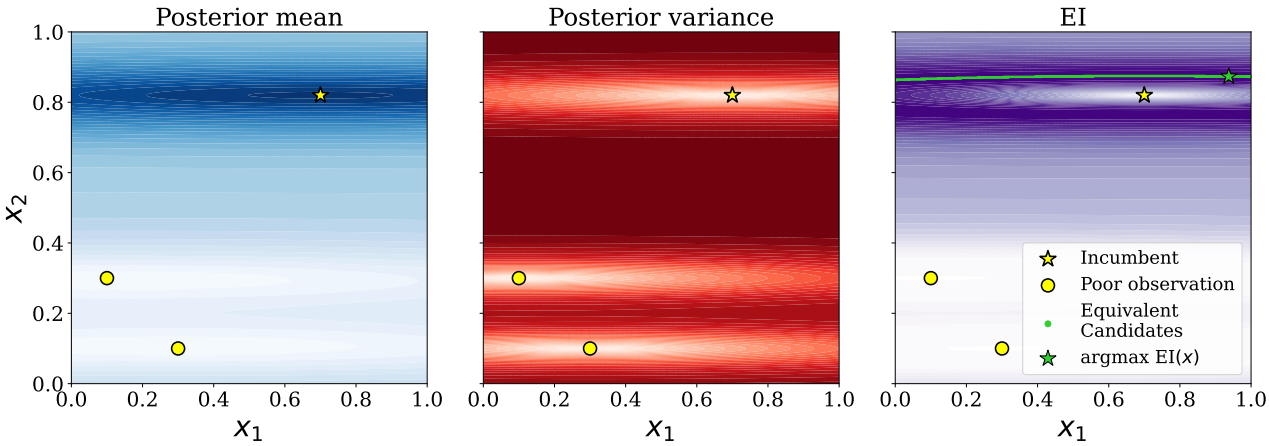


Figure 10. A 2D-visualization of the locality issue for a two-dimensional model of high complexity. The model has a very short lengthscale in x_2 ($\ell_1 = 0.1$), and a long lengthscale in x_1 ($\ell_2 = 1.75$). From left to right, the plots visualize the posterior mean, variance and `EI`, where darker colors mean larger values. Close to the incumbent (yellow star), the EI values are high. The green line in the rightmost plot indicate a region of almost equivalent candidates to the $\arg\max$ (green star) where EI is within a factor 10^{-5} of the $\max$.

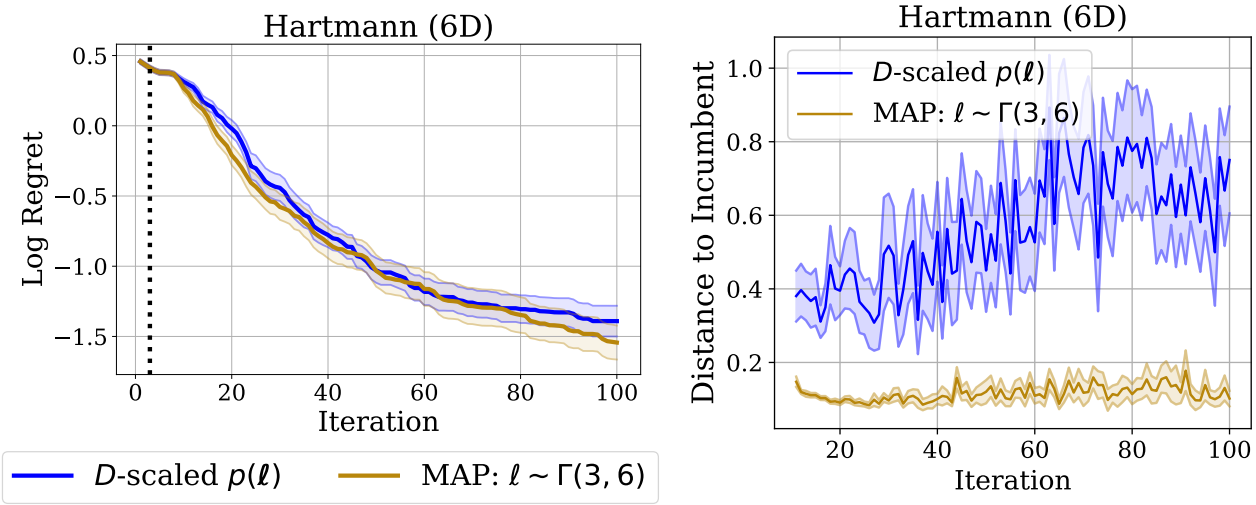


Figure 11. (left) Regret performance a short, conventional lengthscale prior (yellow) and the DSP (blue) across 10 repetitions. Conventional MAP and DSP perform comparably in final performance. (right) Average distance to the incumbent for conventional MAP and DSP per iteration, averaged over 20 repetitions. Conventional MAP exemplifies the locality issue, as it almost exclusively queries in close proximity to the incumbent, whereas DSP displays a balance between exploration and exploitation, as visualized by the wide error bars.

for each method by considering the minimal distance to a preceding data point for each query throughout the optimization round. Fig. 11 visualizes this, and we see that conventional MAP almost never makes an exploratory query after the DoE phase, and takes substantially smaller local steps than our enhancement. While final performance is not negatively impacted, Fig. 11 reveals that conventional MAP is effectively performing an initial random search followed by 140 iterations of local search. While potent, the high complexity of the model effectively disables the global, exploratory component of the BO algorithm.

D. Low-budget Performance

We display the performance of the DSP on the real-world benchmarks from Sec.6. In Fig. 12, we observe that for all

real-world tasks, the DSP provides the strongest initial performance. DSP rapidly improves after initialization, which indicates potent inference ability. Notably, the DSP does not rely on identifying a subset of active variables, yet provides superior performance to the alternatives.

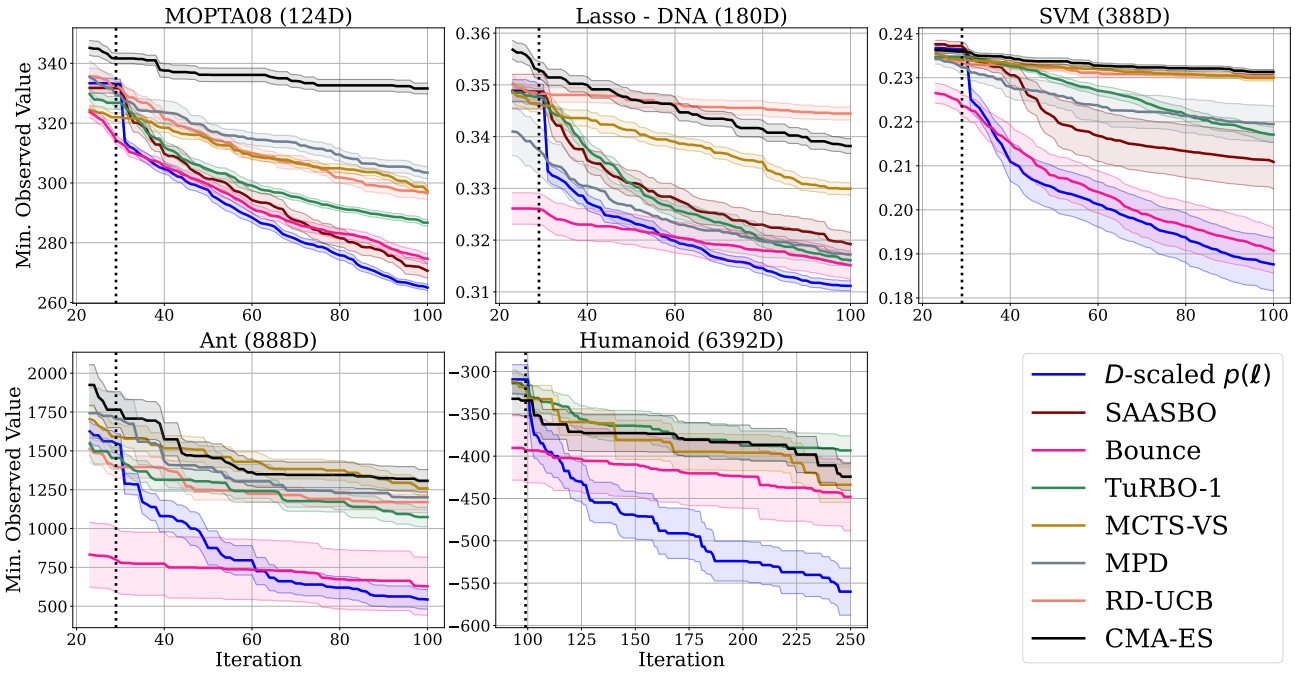


Figure 12. Best observed value of all baselines on five real-world tasks from of various domains across 20 repetitions (10 for SAASBO) and Humanoid. The DSP performs best across all tasks for a low-budget settings, and offers a pronounced initial jump in performance after DoE.

E. Ablation Studies

We ablate the hyperparameters of the enhanced vanilla BO algorithm. Specifically, we provide ablations for both the assumed complexity and the uncertainty of the complexity measure, as well as comparisons to BO with conventional MAP priors (Balandat et al., 2020) of $\ell_i \sim \Gamma(3, 6)$ with high density on low values for the lengthscales, as well as MLE. Lastly, we demonstrate the performance of our enhancement with a learnable signal variance hyperparameter σ_f^2. We evaluate on a subset of the tasks presented in Sec. 6, and include both synthetic, axis-aligned tasks and real-world tasks.

We find that generally, Vanilla BO is highly performant across almost an order of magnitude of complexity, as displayed in Fig. 17, and very robust to changes in uncertainty of the complexity, as shown in Figs. 15 and 16. Assuming too low complexity, however, can occasionally be detrimental to performance, as displayed in Fig 22.

E.1. Conventional MAP and MLE

First, we show the performance of our enhancement to vanilla BO compared to MLE and MAP using the default BoTorch (Balandat et al., 2020) priors ($\ell_i \sim \Gamma(3, 6)$) on the hyperparameters, which match the description given on conventional priors in Sec. 2.2. In the higher-dimensional tasks, MLE hyperparameter fitting is very unstable, due to the large number of hyperparameters in relation to the number of observations. Thus, full repetitions of MLE-estimated hyperparameters are not available for many tasks. We run a stabilized variant of MLE fitting as a complement, where the previous iteration's hyperparameters are in case of a crash during model fitting. This variant achieves competitive performance on many tasks, further demonstrating the capabilities of Vanilla Bayesian optimization fro high-dimensional tasks.

We observe that MAP estimation achieves impressive performance for some tasks, even marginally outperforming our method on a few tasks. We note, however, that conventional MAP severely suffers from the locality issue on all tasks, as it

simply performs a BO-guided local search around the best point found during DoE. We visualize the locality behavior of all three methods in App. C.2.

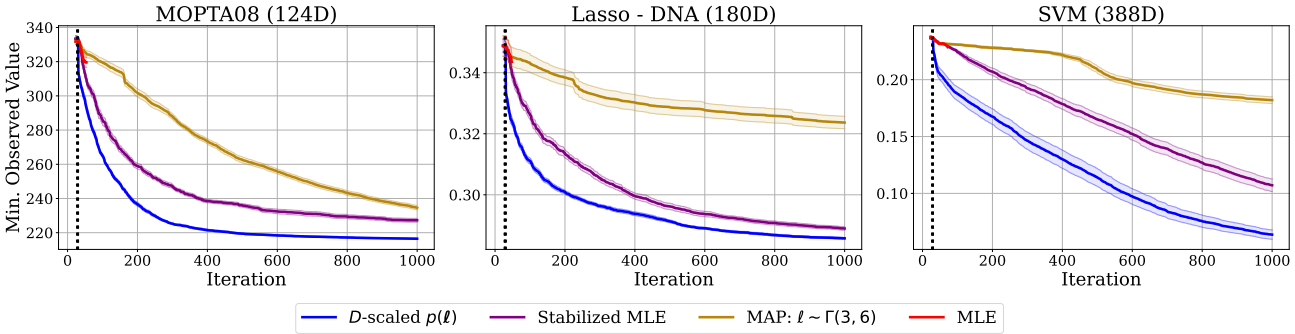


Figure 13. Average log regret of uncertainty ablations on Levy (4D) and Hartmann (6D) synthetic test functions of varying ambient dimensionality for DSP (blue), MLE-estimated hyperparameters (red), stabilized MLE-estimated hyperparameters (purple) and hyperparameters fit through BoTorch default hyperpriors ($\ell_i \sim \Gamma(3, 6)$, green) across 10 repetitions (20 for the default algorithm in blue). $\ell_i \sim \Gamma(3, 6)$ priors perform well on the Hartmann tasks, but searches extremely locally around the best found point from DoE. MLE runs into errors related to fitting instability early on for a majority of all tasks.

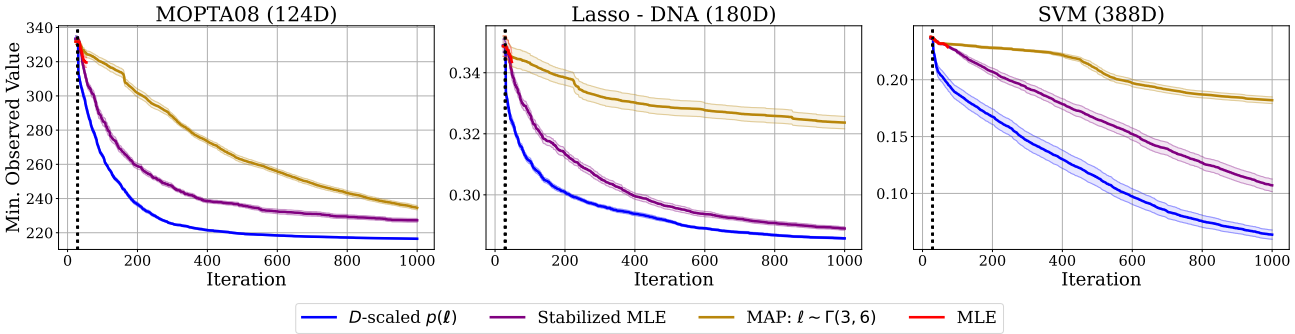


Figure 14. Average minimal observed value on MOPTA, Lasso-DNA and SVM for DSP (blue), MLE-estimated hyperparameters (red), stabilized MLE-estimated hyperparameters (purple) and hyperparameters fit through BoTorch default hyperpriors ($\ell_i \sim \Gamma(3, 6)$, green) across 10 repetitions (20 for the default algorithm in blue). $\ell_i \sim \Gamma(3, 6)$ priors perform well on MOPTA, and MLE runs into errors related to fitting instability early on for all tasks.

E.2. Fixed Complexity - Varied uncertainty

We ablate the uncertainty of the LogNormal prior, keeping the mode fixed. Since the mode of the LogNormal distribution is computed as $\mu - \sigma^2$, we must adjust the μ term when changing σ to keep the mode in place. A larger uncertainty term will increase the ability for Vanilla BO to disregard dimensions that appear unimportant, but also be quicker in identifying dimensions of high importance.

E.3. Varied Complexity - Fixed uncertainty

We ablate the complexity of the method, keeping the σ parameter fixed while varying μ. A larger μ will increase the default lengthscales, suggesting a lower complexity on the problem at hand. We ablate μ by 0.5 in both a positive and negative direction, which corresponds to a factor 1.65 increase and decrease from the default, respectively.

F. Learning the signal variance

In our method in Sec. 5, we propose to not learn the signal variance to safeguard against possible degeneracies related to the active acquisition of data. We ablate the choice here, and demonstrate that, while the performance of a model with a

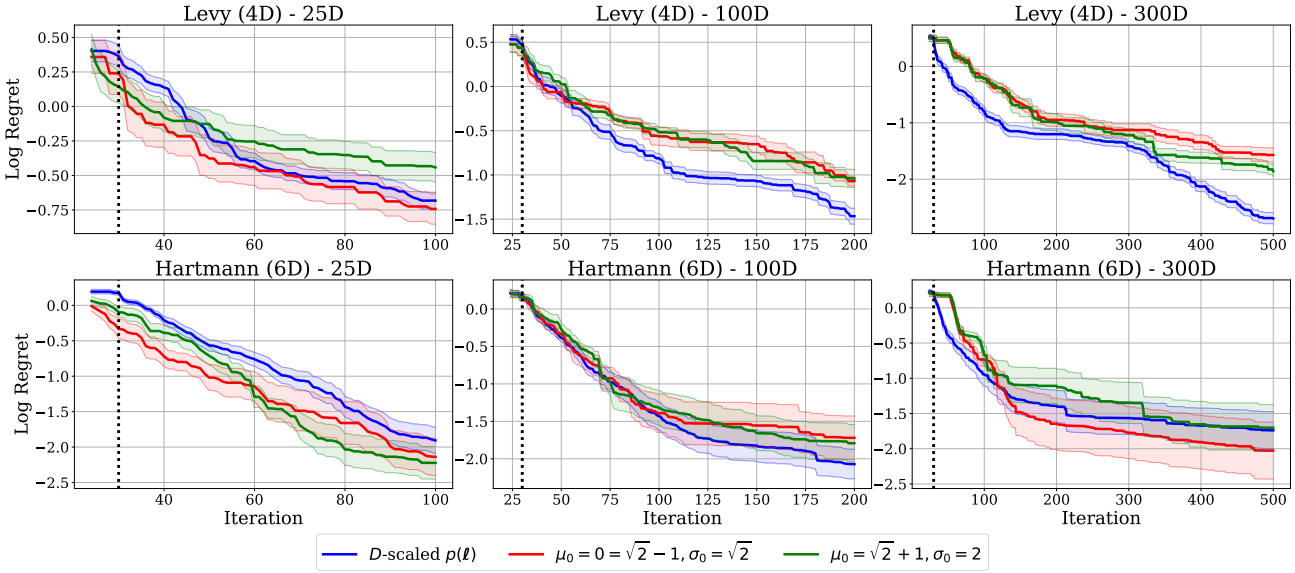


Figure 15. Average log regret of uncertainty ablations on Levy (4D) and Hartmann (6D) synthetic test functions of varying ambient dimensionality across 10 repetitions (20 for the default algorithm in blue).

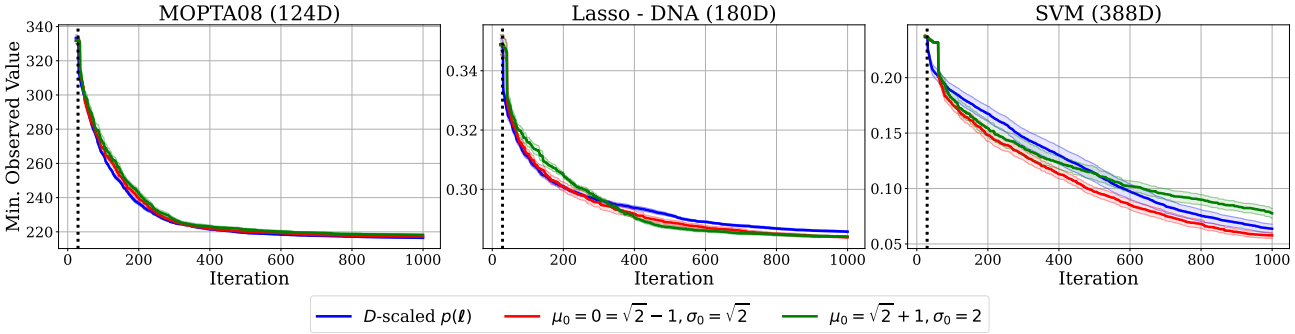


Figure 16. Average minimal observed value on MOPTA, Lasso-DNA and SVM for Vanilla BO with identical modes for the prior, but higher uncertainty (red) and lower uncertainty (green) across 10 repetitions (20 for the default algorithm in blue). Performances are comparable across uncertainty levels, with high uncertainty being marginally worse on SVM.

learned signal variance is marginally less stable, the performance is comparable to one where the signal variance is not learned. In Fig. 19, we display how the outputscale hyperparameter shrinks throughout optimization for models with a learned outputscale. Notably, the fixed outputscale variants (red, blue) do not display constant outputscales due to varying data normalization over time. Moreover, in Fig. 20, we show the exaggerated local search-like behavior that follows from either a learned outputscale or conventional lengthscale prior, as measured by the average distance between the current best point and the upcoming query. The initial samples of high variance and large distance between queries indicates the initial design of experiments via SOBOL sampling.

F.1. Learning σ_f^2

We empirically demonstrate the behavior of the model when learning the signal variance for two tasks, MOPTA08 and SVM, for four types of models throughout the BO loop: DSP with fixed σ_f^2, DSP with learned σ_f^2, and a conventional $\Gamma(3, 6)$ prior with and without learned σ_f^2.

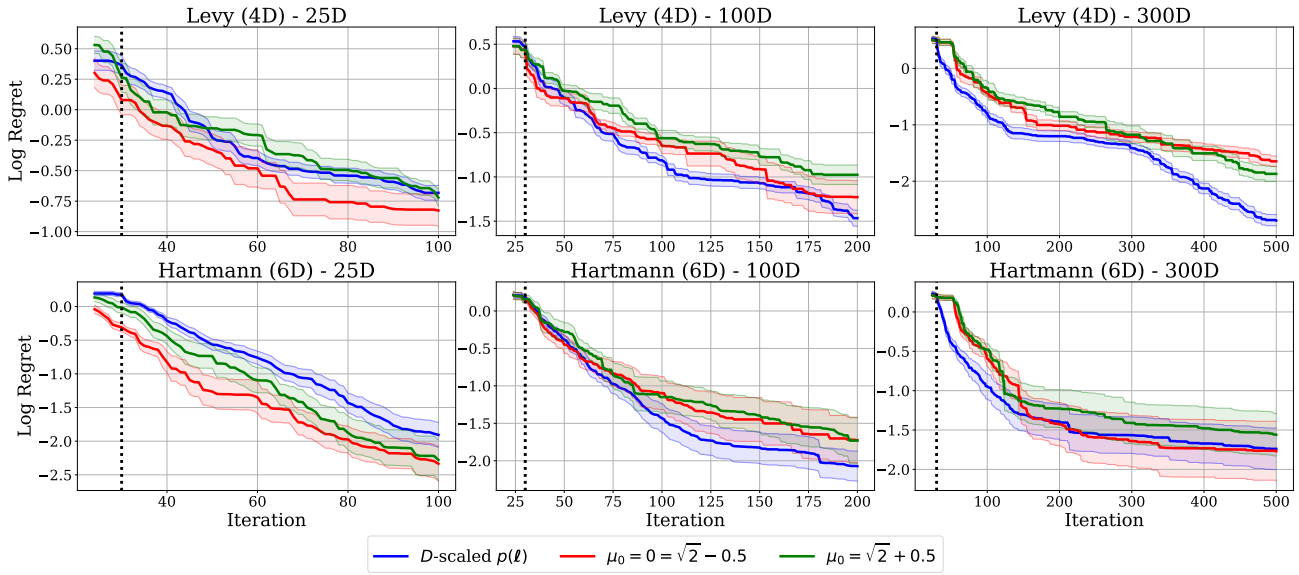


Figure 17. Average log regret of DSP with higher assumed complexity (red) and lower (green) on Levy (4D) and Hartmann (6D) synthetic test functions of varying ambient dimensionality across 10 repetitions (20 for the default algorithm in blue). All instantiations perform comparably, with the low-complexity variant (green) performing slightly worse on average.

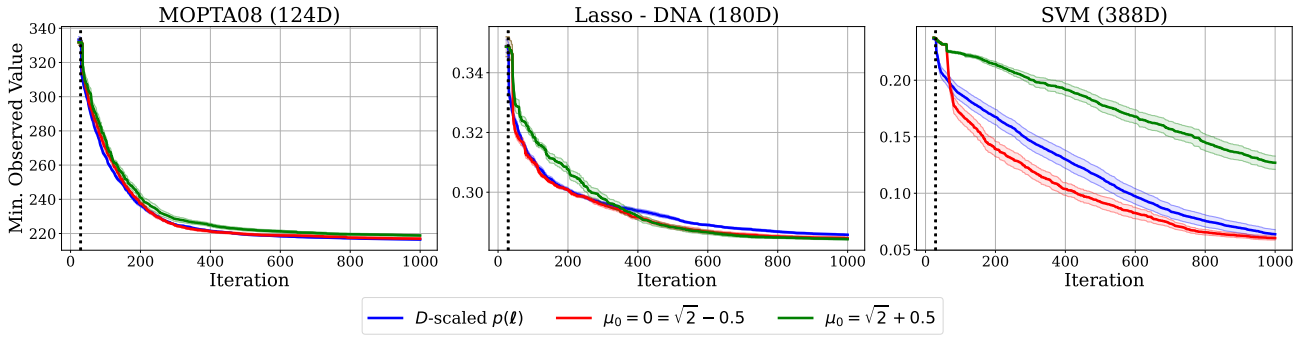


Figure 18. Average minimal observed value on MOPTA, Lasso-DNA and SVM for DSP with higher assumed complexity (red) and lower (green) across 10 repetitions (20 for the default algorithm in blue). Assuming a substantially higher complexity may have detrimental impact on performance, as demonstrated in SVM for the instantiation with the largest μ_0.

F.2. Performance Ablation on Learned σ_f^2

We display the performance of the DSP with and without learned signal variance for a subset of the synthetic and real-world tasks presented in the paper. While the performance is occasionally competitve, the performance is substantially less stable than for the fixed outputscale variant.

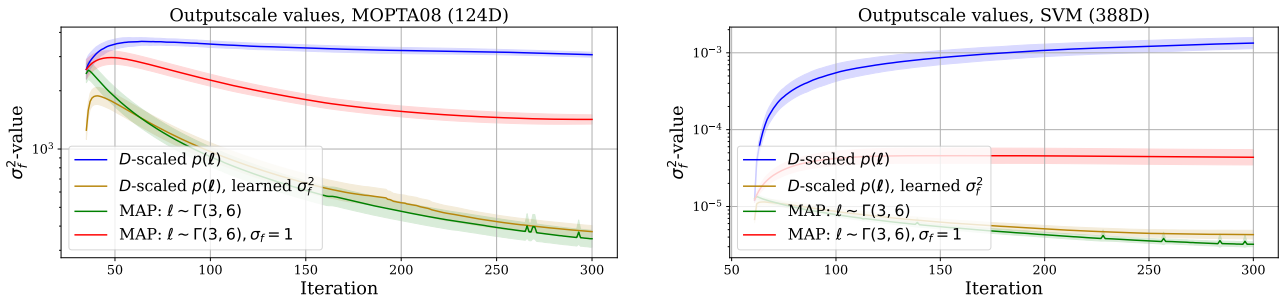


Figure 19. Outputscale shrinkage on MOPTA08 and SVM for BO (mean and one standard deviation) using four types of models employing either the DSP or conventional MAP, and either fixed or learned outputscale. Learned outputscale results in very pronounced shrinkage over time, which leads to increased exploitative behavior.

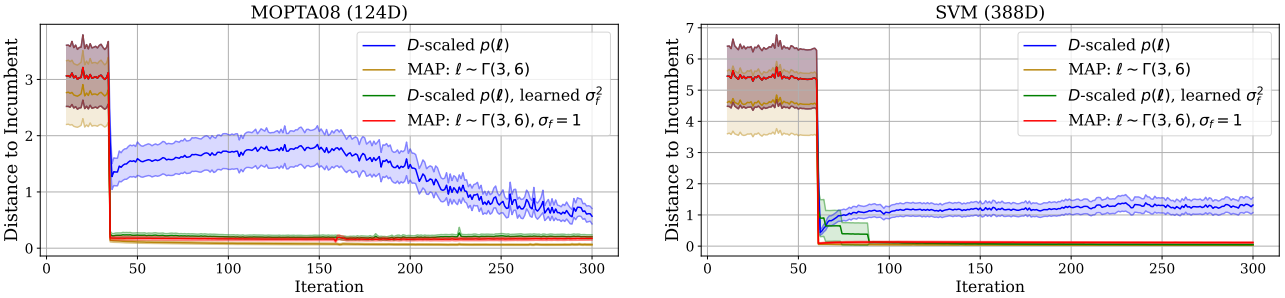


Figure 20. Average distance from current incumbent (mean and 0.25 standard deviation) on MOPTA08 and SVM for BO using four types of models employing either the DSP or conventional MAP, and either fixed or learned outputscale. Learned outputscale or a conventional prior leads to very local optimization, as the algorithm almost exclusively stays within very close proximity to the current best value.

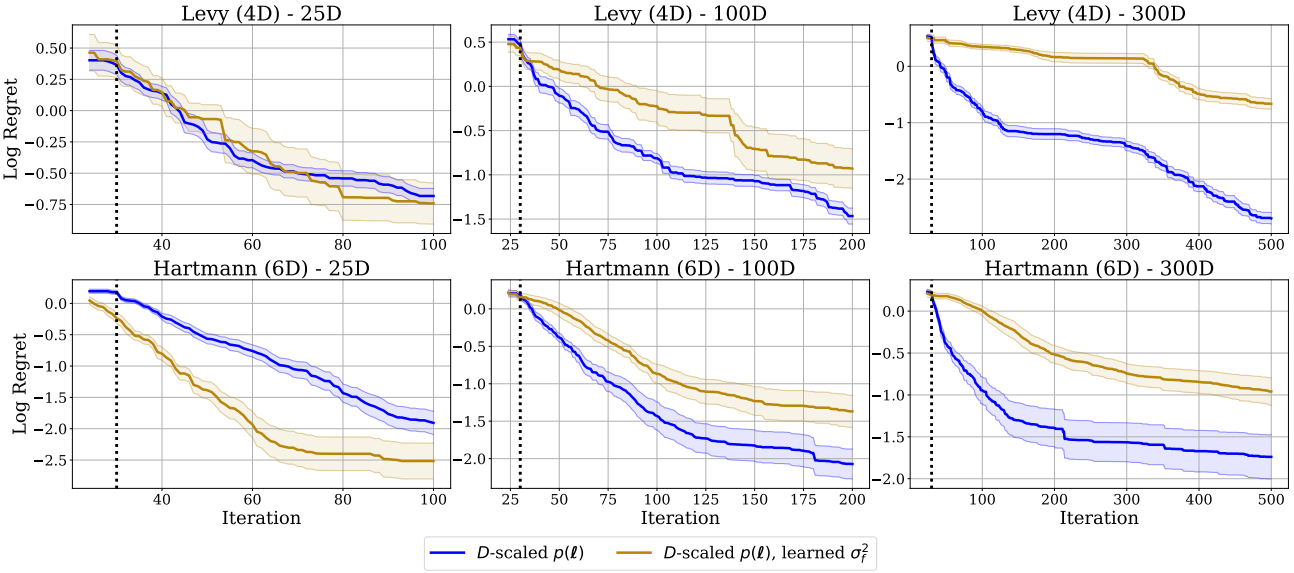


Figure 21. Average log regret of Vanilla BO with a learned signal variance (yellow) and the default algorithm (blue) on Levy (4D) and Hartmann (6D) synthetic test functions of varying ambient dimensionality across 10 repetitions (20 for the default algorithm in blue). The learned signal variance variant performs comparably on some tasks, but displays worse consistency across tasks.

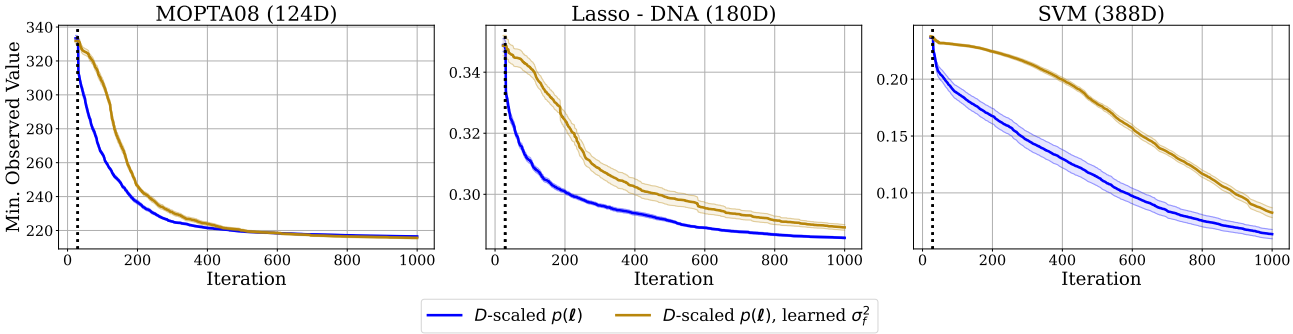


Figure 22. Average minimal observed value of Vanilla BO with a learned signal variance (yellow) and the default algorithm (blue) on MOPTA, Lasso-DNA and SVM across 10 repetitions (20 for the default algorithm in blue). The learned signal variance variant performs comparably on some tasks, but displays worse consistency across tasks.